Nonlinear anisotropic diffusion filters for the numerical approximation of conservation laws

Vom Fachbereich für Mathematik und Informatik
der Technischen Universität Braunschweig
genehmigte Dissertation zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)
von

Thorsten Grahs

Eingereicht am 19. August 2002
Tag der mündlichen Prüfung. 20. Dezember 2002

Referenten: Prof. Dr. Thomas Sonar, TU Braunschweig (Mentor)
Prof. Dr. Hermann Matthies, TU Braunschweig
Prof. Dr. Joachim Weickert, Universität des Saarlandes

Meinen Eltern in Dankbarkeit gewidmet

≪: $\log\cos 41^\circ\, 9'\, 0''$; wo die genau=gleichn letzten 3 Ziffern, sich über die nach oobm & untn anschließend 10 – in Wortn : zehn ! – Werte erstreckn ! – Da bißDe platt, was? ≫/ Sie war nichts weenijer als das. Schprach undeutlich aber fest etwas von <Brot=loosn Künnstn>, von denen <kein Mensch leebm> könnte.

Arno Schmidt "KAFF auch MARE CRISIUM"

Preface

In the last decades the numerical approximation of nonlinear partial differential equations has made a tremendous progress because of the rapid development of computer and information technology. This subject is of particular interest because the analytical knowledge in this area is relatively small. Especially if one compared it to the requirements coming from theory and application in computational fluid dynamics (CFD) and related areas. Therefore, the need for new numerical methods and algorithms increases with the wide appearance of such equations and the increasing number of researchers working on problems arising from this field.

In general, systems of conservation laws are a special set of nonlinear partial differential equations. They constitute very powerful and important models for physical phenomena arising from conservation principles. Important examples are namely the conservation of mass, momentum and energy. They arise in many different topics like fluid mechanics, astrophysics, reactive flows, traffic modelling and several related areas.

From the mathematical point of view, conservation laws are particularly interesting. They tend to develop discontinuous solutions even from smooth initial values because of the nonlinear nature of the equations. This results in a collapse of the classical theory and requires the notion of weak solutions. Therefore, this behaviour leads to important questions concerning the numerical approximation of this class of nonlinear partial differential equations.

One of the central questions emerging in this area is the connection between stability and accuracy of a numerical scheme which approximates discontinuous solutions. For example high-order accurate algorithms using central approximations for derivations are by no means stable. At least in the vicinity of discontinuities they develop oscillations, i.e. the Gibbs phenomenon occurs. Even more sophisticated high-order schemes using limiter functions or reconstruction algorithms reduce to first order accurate approximations at discontinuities. This order reduction in accuracy is due to the necessary artificial dissipation which stabilises the approximative solutions.

Therefore, numerical schemes for hyperbolic conservation laws have to serve two different purposes – high accuracy and sharp shock resolutions – which are in a way contrary to each other. One has to find alternatives in order to fulfil both requirements in a well-balanced way. Hence, there is a strong need for nonlinear filter algorithms which reduce the oscillatory behaviour of high order schemes. In order to maintain the accuracy one has to relax monotonicity requirements and admit small variations in the

vicinity of shocks. The question is how to accomplish the construction of such discrete filter terms which stabilise the numerical approximation without the loss of accuracy.

Over the past years a theory of anisotropic diffusion filters was developed in an area which has at a first glance nothing in common with the area of CFD or numerical approximations of conservation laws – image processing and computer vision. There partial differential based filters led to an entirely new field where they are used for structure detection, edge enhancement and noise reduction.

Looking at this mathematical subject in more detail several relations between both fields, namely CFD and image processing can be explored. For example an oscillatory discontinuity can be regarded as a noisy edge and the question of sharp shock resolution may be related to edge enhancement. Consequently there are already several different approaches in the area of image processing which are based on ideas from the numerical treatment of conservation laws.

There might be a hint to follow the path in the opposite direction and put the question vice versa: How can ideas and methods from image processing be used for the numerical solution of conservation laws. This thesis addresses this problem. The idea of integrating anisotropic diffusion models into numerical schemes for conservation laws is entirely new. These filter schemes are related to the Perona-Malik model which is a nonlinear data-dependent diffusion model for edge enhancement. The algorithms developed by Weickert can be regarded as an genuinely multidimensional or anisotropic extension of this filter, since they introduce direction dependent diffusion, based on informations gained from structure detection tools.

Since these diffusion models are genuinely multidimensional the approach is to incorporate them into multidimensional algorithms. A one-dimensional approach, which leads to a kind of Perona-Malik diffusion model was recently proposed by Wei [114]. This results in an approach with is similar to the Essentially-Non-Oscillatory (ENO) or Weighted-ENO (WENO) methods.

On the other hand, a naive integration of this new filter algorithms will lead to unsatisfying results. Numerical tests show that it is necessary to take into account additional information gained from the original equations in order to steer the developed filters in a most suitable way. For our purpose it seems to be successful to use entropy conditions as additional information for the design of the nonlinear anisotropic diffusion models. These conditions are strongly related to conservation laws. This development is presented in the following. In addition, a first application to systems of conservation laws, namely the Euler equations of gas dynamic is presented. This application shows that the use of image processing tools even for systems of conservation laws is possible.

Organisation of the thesis

This thesis will start with the presentation of the theoretical framework of conservation laws. A survey over the fundamental problems, ideas and solution approaches is given for scalar equations as well as for systems of conservation laws.

In the second chapter notions and theoretical results as well as important algorithms for the numerical solution of scalar conservation laws will be introduced. The extension to systems is in general straight forward and can be found in the cited textbooks. Then we introduce numerical entropy inequalities which will serve later as a detection tool for the developed algorithms. In this field several tasks are combined, e.g. the choice of the numerical approximation of the entropy flux function and the consistent foundation of this choice. This new derivation developed in this thesis leads to alternative formulations for classical results concerning entropy satisfying schemes and their limits.

Chapter 3 gives a condensed overview over diffusion filters in the context of image processing and computer vision. This survey considers linear and nonlinear models. On the basis of the Perona-Malik model the behaviour of nonlinear diffusion models based on forward- and backward-diffusion will be analysed. This is followed by the introduction to anisotropic filters due to the developments by Weickert. They are based on a nonlinear diffusion model which is founded on a data analysis tool, called the structure tensor. This tensor puts a strong data dependence into these filters.

The next chapter is the heart of this thesis, since it is dealing with the development and integration of nonlinear anisotropic diffusion models into high-order accurate oscillatory schemes. We present several approaches concerning the construction of such artificial dissipation models for scalar conservation laws. Since this approach is entirely new, we start with a straight forward approach. Then we present sophisticated data analysis tools for structure detection. These tools are based on features arising in the theory of numerical approximations of conservation laws – namely cell entropy inequalities – developed in the second chapter and positivity requirements.

Chapter 5 transfers ideas developed in the foregoing section to systems of conservation laws. Here, the Euler equations of gas dynamics are chosen, since they are the most prominent system in this context. The algorithm demonstrates that it is in principle possible to apply nonlinear anisotropic diffusion filters to systems of conservation laws.

Chapter 6 is devoted to the presentation of numerical results gained with the new developed schemes in both foregoing chapters. A test case is described for the scalar algorithms as well as the system approach. In the following the test case is applied to the schemes.

Finally, the thesis concludes with a short summary and an outline of future perspectives and ongoing research concerning the topic of nonlinear anisotropic diffusion filters.

Acknowledgements

This work was written with the support of a graduate scholarship of the University of Hamburg at the Institute for Applied Mathematics and at the Institute for Analysis at the Braunschweig University of Technology where I was working as a research assistant. During this time I was partly supported by a scholarship from the Richard-Winter-Foundation for science and research.

During the years working on this subject I have been helped and influenced by many people, and it is a pleasure to use the opportunity here to express my gratitude to them.

At first, I have to thank my supervisor Prof. Dr. Thomas Sonar, who made me familiar with the subject of anisotropic diffusion filters and supported my work with permanent interest, advice and discussions. Not only creating a gentle atmosphere of kindness and humanity, he also allowed great freedom in work and research. For all of this I like to express deep gratitude.

I also like to thank Prof. Hermann Matthies, Ph. D., Head of the Institute of Scientific Computing at the Braunschweig University of Technology for his kind willingness to referee this thesis and the exam.

Further on I would like to thank Prof. Dr. Joachim Weickert from the Institute for Applied Mathematics, University of Saarbrücken, for fruitful discussions concerning the field of anisotropic diffusion in image processing and his continuous encouragement. Likewise, I am grateful for his immediate preparedness to act as a further referee.

I like to thank Prof. Dr. Remi Abgrall from the Institute of Applied Mathematics at the University Bordeaux for the invitation to spend the Summer 2000 at the Centre international de recherches en mathématiques (CIRM) in Marseille. Beside many other things I learned from him that congeniality needs confusion.

Since this work was prepared in Hamburg and Braunschweig, I like to thank all my colleagues and collaborators at the respective institutes. For the Hamburg years I like to emphasise a special thanks to Dr. Lars Hoffmann and Priv. Doz. Dr. Andreas Meister for helpful discussions, a comfortable atmosphere and necessary coffee breaks.

For the last years in Braunschweig, I like to thank all the members at the Institute of Analysis, and especially the members of the group Sonar, Mrs. Dorothea Agthe, Dipl.-Math. Andrea Bürgel, Dipl.-Math. Stefanie Schmidt and Dr. Ingo Thomas for a relaxed time and continuous support.

A special thanks to Prof. Dr. Joachim Wirths for being occasionally his research assistant a long time after being his student. In addition I learned from him that it is easier to handle an audience of five hundred engineers than herding a sack full of flees (German saying).

A deep and special thanks to my parents, Irmgard and Günter Grahs for their contin-

uous encouragement and support during the past years.

Last, but not least, I like to thank Dr. Jens-Peter Bode, Dipl.-Math. Andrea Bürgel, Dipl.-Ing. Gesche Hagemann, Priv. Doz. Dr. Andreas Meister, Dipl.-Psych. Josephine Scheithauer Dr. Nils Scheithauer, Dr. Ingo Thomas, Dr. Stephan Venzke, Mrs. Anja Venzke, lawyer at local court, and Dipl.-Ing. Stephan Will for reading the manuscript carefully and helping in all kinds of difficulties concerning the English language. To them, all my other friends and the products of several local german breweries and french vineyards I like to say:
"*Thanks for being there!*"

Braunschweig, August 2002 Thorsten Grahs

Contents

The universe is a gigantic system of reflexes produced by shocks.

Bernard Shaw
("The black girl in search of God")

1 Conservation laws

As introduction we give a condensed overview of the theoretical results and notions for systems of conservation laws in several space dimensions. The governing equations and their major properties will be presented.

Then we make some specifications and examples of hyperbolic conservation laws. We also introduce the solution theory following [21, 31, 32, 95, 96, 113].

1.1 Fundamental principles

In this thesis a particular class of nonlinear partial differential equations is considered, namely those of hyperbolic type. The hyperbolic nature reflects the fact that these equations to model transport or advection phenomena. In contrast to parabolic or elliptic partial differential equations which describe diffusion or equilibrium respectively, those of hyperbolic type model physical systems dominated by advection, like wave or flow phenomena. This is easily demonstrated by the most simple model for this class, the one dimensional wave equation. It is the mathematical description of the spreading of a wave-like pattern.

The mathematical models describing flow phenomena in the physical world are fully or partly hyperbolic. The Navier-Stokes equations, describing the movement of fluids under influence of body forces and friction, are partly hyperbolic due to the convective terms dominating these equations in general. However, they are also parabolic since friction and hence dissipation play an important role particularly in boundary layers.

The Euler equations are the best investigated mathematical model for a hyperbolic system of conservation laws. They describe the flow of compressible gases or liquids at high pressure where viscous effects can be neglected. From the mathematical point of view they are interesting because of their nonlinear behaviour, which implies the development of discontinuous solution from smooth initial data in finite time.

The notion conservation laws refers to the fact, that they summarise some principal physical laws, namely the conservation of mass, Newton's Second Law and the conservation of energy. The formulation in the form of conservation laws describes the conservation of the considered quantities.

Like Majda [77] pointed out, a conservation law arise from modelling physical processes in three steps:

i) It reveals a physical balance law derived from p physical quantities $u_1, \ldots, u_p$ combined in a vector $\underline{u} = (u_1, \ldots, u_p)^T$ with $\underline{u}(t, \underline{x})$ mapping from the space $(t, \underline{x}) = (t, x_1, \ldots, x_d)^T$ spanned by one time and d space dimension into an open subset $\mathcal{S} \subset \mathbb{R}^p$, the so-called state space. The state space $\mathcal{S}$ arises from the fact, that several physical quantities, e.g. pressure or density, are nonnegative functions.

ii) The flux functions appearing in the balance law are idealised by prescribed nonlinear functions $f_j(\underline{u})$, mapping $\mathcal{S}$ into $\mathbb{R}^p$. Since we neither consider source terms $S(\underline{u}, t, \underline{x})$ like external body forces or heat sources, nor microscopic feature like diffusion and dissipation, the balanced forces are conserved. This gives rise to the conservation law.

iii) A generalised version of the principle of virtual work is applied [5].

The conservation of mass shall be treated as an example for this abstract formulation of the origin of a physical conservation law.

Conservation of mass

We consider a scalar quantity $\rho(t, \underline{x}) : \mathbb{R}^{d+1} \to \mathbb{R}$ which describes the state of the quantity in a point $\underline{x} \in \mathbb{R}^d$ at time t. We can think of ρ as the density of a streaming fluid, but this is not necessary. If we are interested in the development of the quantity, it is reasonable to ask for the change in time. For that we consider a volume V. The quantity accumulated in this volume at time t is

$$Q(t, \mathsf{V}) \quad = \quad \int_{\mathsf{V}} \rho(t, \underline{x}) \, d\underline{x}. \tag{1.1}$$

If we assume that volume V is impermeable and that mass is neither created nor destroyed, the mass located in V can only change by exchange through the volume faces[1].

Assuming that the velocity of the gas at the point $\underline{x}$ at time t is given by $\underline{v}(t, \underline{x})$ the flow rate or flux of the gas is given by $\rho(t, \underline{x})\underline{v}(t, \underline{x})$. Since we are interested in the change of the mass in the volume V in time, we have to examine the derivative with respect to time for (1.1). Due to our consideration above – or the physical principle we like to reveal – this is balanced by the flow through the surface of the volume, i.e.

$$\frac{d}{dt} \int_{\mathsf{V}} \rho(t, \underline{x}) \, d\underline{x} \quad = \quad - \int_{\partial \mathsf{V}} \rho(t, \underline{x}) \underline{v}(t, \underline{x}) \cdot \underline{n} \, d\underline{\sigma}, \tag{1.2}$$

where $\partial \mathsf{V}$ is the surface of the volume V. If the density is sufficiently smooth, we may interchange differentiation and integration and, applying the Gauß or divergence Theorem on the right-hand side, we derive

$$\int_{\mathsf{V}} \left[\partial_t \rho(t, \underline{x}) + \nabla \cdot (\rho(t, \underline{x}) \underline{v}(t, \underline{x})) \right] \, d\mathsf{V} \quad = \quad 0. \tag{1.3}$$

[1] Here, we already start to introduce a physical principle into our mathematical model. One can easily guess which kind of principle we like to deduce.

Since this must hold for an arbitrary control volume (1.3) holds pointwise and the integrand has to be identically zero, i.e.

$$\partial_t \rho(t,\underline{x}) + \nabla \cdot (\rho(t,\underline{x})\underline{v}(t,\underline{x})) \quad = \quad 0. \tag{1.4}$$

This is the divergence form of the conservation of mass, while (1.3) is called the integral form. Hence, merely on the assumptions of neither creating nor destroying mass we have derived the mathematical model for conservation of mass. Later we will see that all equations which model conservation properties have the form

$$\partial_t \underline{u} + \sum_{j=1}^{d} \partial_{x_j} \underline{f}_j(\underline{u}) = 0. \tag{1.5}$$

We call this type of systems of partial differential equations systems of conservation laws. Later we will give the adequate mathematical definition for this set of equations.

1.2 Scalar conservation laws

As we have seen the hyperbolic nature of the equations considered is strongly related to the advection or wave spreading, which is modelled in this class of partial differential equations. The simplest model for this equation type is the scalar wave equation with constant speed. This equation describes the transport of a scalar quantity u depending on the direction and the velocity. This equation will be considered in a more detailed way by examining the behaviour of the solution while changing to a nonlinear form.

Transport phenomena

In our derivation of the principle of conservation of mass, we assumed that the change of mass in the control volume V only takes place by flow through the cell faces. This means that our model does not contain sinks or sources. Furthermore, if we neglect viscous phenomena, body forces etc., we have only transport or advection between the cells. If we assume the simplest form of this process, which means taking the velocity vector as constant, i.e. $\underline{v}(\underline{x},t) = \underline{v}_{\mathsf{const}} = \mathsf{constant}$, we obtain

$$\partial_t u + \underline{v}_{\mathsf{const}} \nabla u \quad = \quad 0. \tag{1.6}$$

As one can easily see the development in time is balanced by a drift or transport with velocity $-\underline{v}_{\mathsf{const}}$. So the change of $u(t,\underline{x})$ depends on the scalar product $\langle \underline{v}_{\mathsf{const}}, \nabla u \rangle$. For simplicity we assume that initial conditions only consist of constant states u^- and u^+ separated by a discontinuity Σ. The analytic solution of the Cauchy problem (1.6) with initial condition

$$u(0,\underline{x}) \quad = \quad u_0(\underline{x}) \tag{1.7}$$

is simply

$$u(t,\underline{x}) \quad = \quad u_0(\underline{x} - \underline{v}_{\mathsf{const}} t). \tag{1.8}$$

Here, one immediately sees that the initial data propagate with velocity $\|\underline{v}_{\mathsf{const}}\|$ in space-time along the rays $\underline{x} - \underline{v}_{\mathsf{const}} t = \underline{x}_0$.

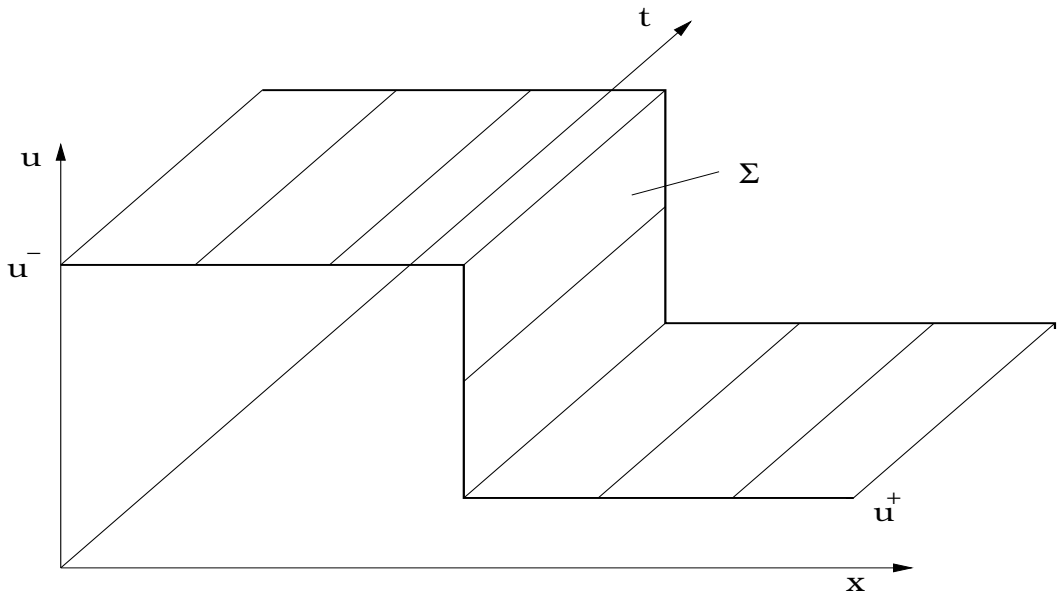

Figure 1.1: Initial condition u^- and u^+

Characteristic curves

Considering the solution (1.8) of the problem (1.6), (1.7) one gets the impression that these lines play an important role for the development of the solution $u(\underline{x}, t)$ in time. And indeed considering a one-dimensional example one can see, that the initial data (1.7) are translated to the right (resp. to the left) for $v_{\mathsf{const}} > 0$ (resp. $v_{\mathsf{const}} < 0$) (see Figure 1.1).

Looking more closely at the solution (1.8) and drawing it at time t_0 and t_1 into a space-time diagram, we clearly see the profile of the initial data is transported along these rays. They are called characteristic curves and are defined as the integral curves of the differential equation

$$\frac{d\underline{x}}{dt} = \underline{v}_{\mathsf{const}}. \tag{1.9}$$

If we examine the change of the solution along theses curves we see that

$$\begin{aligned} \frac{d}{dt}u(t, \underline{x}(t)) &= \partial_t u(t, \underline{x}) + \nabla u(t, \underline{x}) \frac{d\underline{x}}{dt} \\ &= \partial_t u + \langle v_{\mathsf{const}}, \nabla u \rangle \\ &= 0. \end{aligned}$$

This shows that the solution u is constant along these characteristics. Since we have a linear equation and $\underline{v} = \underline{v}_{\mathsf{const}}$ is constant, the characteristic curves are straight lines, which corresponds to linear advection or linear transport.

Nonlinear equations

In the first section we stated that a general conservation law in several space dimensions has the form (1.5). If we restrict ourselves for the moment to scalar conservation laws in several

space dimensions, i.e. $p = 1$, we obtain

$$\begin{aligned} \partial_t u + \langle \nabla, \underline{f}(u) \rangle &= 0, \\ u(0, \underline{x}) &= u_0(\underline{x}), \end{aligned} \tag{1.10}$$

with $\underline{f}(u) = (f_1(u), \ldots, f_d(u))^T : \mathbb{R} \to \mathbb{R}^d$.

Remark 1.1
It is easy to see, that also the linear advection equation (1.6) can be written in this general form, which is called conservative formulation.

With (1.7) we can formulate the Cauchy problem for this equation. If we assume u is a classical solution of (1.10) we can carry out the differentiation of the flux vector $\underline{f}(u)$ and derive the nonconservative form of (1.10), i.e.

$$\partial_t u + \langle \underline{f}'(u), \nabla u \rangle = 0. \tag{1.11}$$

With $\underline{f}'(u) = \underline{a}(u)$ we find a form similar to (1.6) now with the nonlinear velocity $\underline{a}(u)$. Thus, the characteristics now have a more general form than (1.9) and we see that the change in time is balanced by the derivative of the flux vector with respect to u.

(1.11) can be rewritten considering the physical space as space-time-continuum with $x_0 := t$. Thus, we write

$$\begin{aligned} \partial_t u + \langle \underline{f}'(u), \nabla u \rangle &= \partial_{x_0} u + \sum_{j=1}^{d} f_j'(u) \partial_{x_j} u \\ &= [1, f_1'(u), \ldots, f_d'(u)][\partial_{x_0} u, \partial_{x_1} u, \ldots, \partial_{x_d} u]^T \\ &=: \langle \underline{a}^{\mathsf{t}}(u), \nabla^{\mathsf{t}} u \rangle. \end{aligned}$$

Hence, (1.11) can be viewed as the derivation of the conserved quantity u in the direction of the vector $\underline{a}^{\mathsf{t}}(u)$. This form can be interpreted in three different cases which lead to the same conclusion (see [17],[83]):

i) Find a hyperplane $S^{\mathsf{t}} \in \mathbb{R}_+ \times \mathbb{R}^d$ such that the solution u in a neighbourhood of S^{t} is uniquely determined, i.e. $\underline{z}^{\mathsf{t}}$ is normal to S^{t}.

ii) Find a curve $\underline{z}^{\mathsf{t}}$ such that the solutions u on both sides, i.e. in direction normal to the curve, are independent from each other. This is equivalent to the prohibition of carrying out differentiation in the normal direction $\underline{n}^{\mathsf{t}}$.

iii) Find a curve $\underline{z}^{\mathsf{t}} : \mathbb{R} \to \mathbb{R}_+ \times \mathbb{R}^d$ in space-time such that $\underline{a}^{\mathsf{t}}(u)$ is in every point tangential to $\underline{z}^{\mathsf{t}}$. Hence, this is equivalent to carrying out the differentiation only in the direction $\underline{a}^{\mathsf{t}}(u)$.

Following the latter Ansatz we define the curve $\underline{z}^{\mathsf{t}}$ as

$$\underline{z}^{\mathsf{t}}(s) = \begin{bmatrix} t(s) \\ \underline{x}(s) \end{bmatrix}.$$

The derivative $\underline{z}^{\mathsf{t}\prime}$ is tangent to $\underline{z}^{\mathsf{t}}$ which means that $\underline{z}^{\mathsf{t}}$ coincides with $\underline{a}^{\mathsf{t}}$, i.e.

$$\frac{d}{ds}\underline{z}^{\mathsf{t}}(s) \;=\; \underline{a}^{\mathsf{t}}(u(\underline{z}^{\mathsf{t}}(s))), \qquad \underline{z}^{\mathsf{t}}(0) \;=\; \begin{bmatrix} 0 \\ \underline{x}_0 \end{bmatrix},$$

or component-wise

$$\begin{aligned} \tfrac{d}{ds}t(s) &= 1, & t(0) &= 0, \\ \tfrac{d}{ds}\underline{x}_j(s) &= f_j'(u(t(s),\underline{x}(s))), & \underline{x}_j(0) &= \underline{x}_{0_j}. \end{aligned}$$

This system of ODEs is solved by

$$\begin{aligned} t(s) &= s = t, \\ x_j(s) &= x_{0_j} + \int_0^t f_j'(u(s,\underline{x}(s)))\,ds. \end{aligned} \tag{1.12}$$

This leads to the following

Proposition 1.2
Assume that u is a classical solution of (1.10). Then the characteristic curves are straight lines along which the solution is constant.

Proof Again we only have to consider the change of u along a characteristic line (1.12). This exists at least for a short time interval $[0, t_0[$ and we get

$$\begin{aligned} \frac{d}{dt}u(t,\underline{x}(t)) &= \frac{\partial u(t,\underline{x}(t))}{\partial t}\underbrace{\frac{dt}{dt}}_{1} + \sum_{j=1}^{d}\frac{\partial u(t,\underline{x}(t))}{\partial x_j}\underbrace{\frac{dx_j(t)}{dt}}_{f_j'} \\ &= \partial_t u + \sum_{j=1}^{d} f_j'\partial_{x_j}u \\ &= 0, \end{aligned} \tag{1.13}$$

so u is constant along the characteristic curve. From (1.13) follows that the characteristic curves are straight lines with constant slopes depending on the initial data. ∎

Remark 1.3
(1.13) depends strongly on the fact that we start from a conservation law, i.e. the change in time is balanced by the flux through the boundary. If source terms are involved, one clearly sees that the quantity u is not constant along the characteristic curves which are obviously no longer straight lines.

Nonuniqueness

At this point the question naturally arises what happens to a conservation law with nonlinear flux function like (1.10) if characteristic lines intersect. In the linear case (1.6), this problem

does not arise, since the slopes of the characteristic curves are constant, i.e. independent of the initial condition. As we have seen before this is not true for the nonlinear case.

Since the solution u is constant along a characteristic we may get multivalued solutions if characteristic lines cross each other. However, this contradicts our notion of a function.

For simplicity let us restrict to the one-dimensional case and assume that the flux function f satisfies $f'(u_0(x_1)) > f'(u_0(x_2))$ with $f'(u_0(x_1)) \neq 0, f'(u_0(x_2)) \neq 0$ for two points $x_1 < x_2$, i.e.

$$m_1 = \frac{1}{f'(u_0(x_1))} < \frac{1}{f'(u_0(x_2))} = m_2.$$

Since the slope m_2 corresponding to the characteristic C_2 is steeper than slope m_1 of the characteristic C_1, they necessarily intersect at some point P at time

$$t = \frac{x_2 - x_1}{f'(u_0(x_1)) - f'(u_0(x_2))}.$$

Since u can not take both values $u_0(x_1)$ and $u_0(x_2)$, a discontinuity has to arise at the point P. The solution *breaks* and a shock forms. Note that we have not assumed special properties

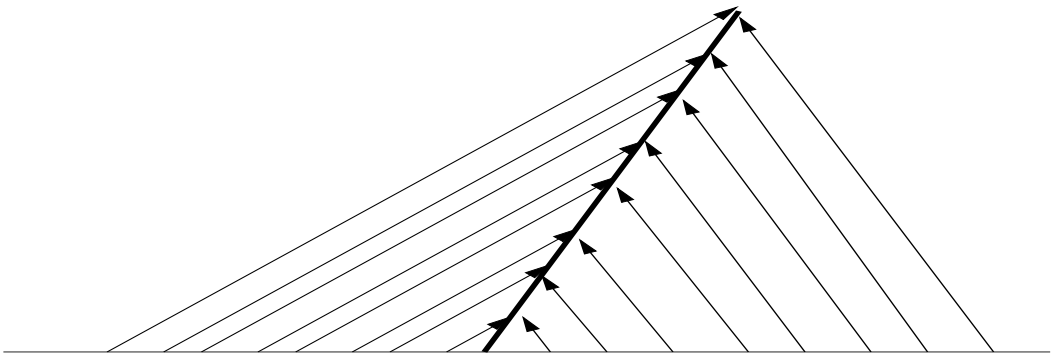

Figure 1.2: Characteristic lines for a shock wave

of u and f, so the solution is independent of the smoothness of both functions. This behaviour is very special for our class of equations: the possible development of discontinuous solutions from smooth initial data. Here the notion of classical solution fails and a new concept is needed. As we will see in the next section this dilemma is solved by the notion of weak solutions.

Weak solutions

The above considerations have clearly shown, that classical solutions are not sufficient to resolve (1.10). Therfore, we have to consider weak solutions, which means solutions in the sense of distributions.

The formulation of a weak solution for (1.10) that does not require differentiability starts

from the integral form of the conservation law, i.e.

$$\int_{t_1}^{t_2}\int_{\Omega}\left[\partial_t u(t,\underline{x})+\nabla\cdot\underline{f}(u(t,\underline{x}))\right]\,d\underline{x}\,dt \;=\; 0.$$

Multiplication with a test function $\phi\in C_0^1([0,+\infty[\times\mathbb{R})$ and carrying out the differentiation to the test function by integration by parts one obtains

$$\begin{aligned}0 &= \int_0^{\infty}\int_{\Omega}\left[\partial_t u+\nabla\cdot\underline{f}(u)\right]\phi\,d\underline{x}\,dt\\ &= -\int_0^{\infty}\int_{\Omega}\left[u\partial_t\phi+\langle\underline{f}(u),\nabla\phi\rangle\right]\,d\underline{x}\,dt-\int_0^{\infty}\int_{\partial\Omega}\langle\underline{f}(u),\underline{n}\rangle\phi\,d\underline{o}\,dt.\end{aligned}$$

Here $\underline{n}$ is the outer normal of $\partial\Omega$. Since the test functions have compact support, i.e. vanish on the boundary $\partial\Omega$, the surface integral vanishes and we derive

$$\int_0^{\infty}\int_{\Omega}\left[u\partial_t\phi+\langle f(u),\nabla\phi\rangle\right]\,d\underline{x}\,dt+\int_{\Omega}u_0(\underline{x})\phi(0,\underline{x})\,d\underline{x} \;=\; 0. \tag{1.14}$$

Thereby, one sees that it is enough to consider u as a measurable function such that $f(u)$ is defined pointwise.

Definition 1.4
The function u is called a weak solution of the Cauchy problem (1.10) if $u, f(u)\in L^1_{\mathsf{loc}}$ and (1.14) holds for all test functions $\phi\in C_0^1(\mathbb{R}\times\mathbb{R}^d)$.

Remark 1.5
Not surprisingly, a classical solution of the Cauchy problem (1.10) is also a weak solution of the problem. On the other hand, every distributional solution is a classical solution of (1.10) in any domain where u is C^1.

For a detailed discussion on weak and measure-valued solutions for conservation laws see [79].

Unfortunately, weak solutions are by no means unique, i.e. the propagation velocity at which discontinuities propagate is not necessarily uniquely determined. Furthermore, not every discontinuity is admissible. A necessary condition to the jump discontinuity is given by the following condition.

The Rankine-Hugoniot condition

We consider u as a piecewise C^1-function, which means in this context, that there exists a smooth orientable surface Σ in the (t,$\underline{\mathrm{x}}$)-space which separates the domain Ω in two connected components Ω^+ and Ω^-, where u is a C^1 function and across the surface u has a jump

discontinuity. u^+ and u^- are the limits of u on the respective sides of Σ (see Figure (1.3)), i.e.

$$u(\underline{x},t) \;=\; \begin{cases} u^+(t,\underline{x}) & , \quad \lim\limits_{\varepsilon\to 0} u((t,\underline{x})+\varepsilon\underline{n}) \\ u^-(t,\underline{x}) & , \quad \lim\limits_{\varepsilon\to 0} u((t,\underline{x})-\varepsilon\underline{n}) \end{cases} .$$

where $\underline{n} = (n_t, n_{x_1}, \ldots, n_{x_d})^T$ is the unity vector normal to the surface Σ, chosen to point

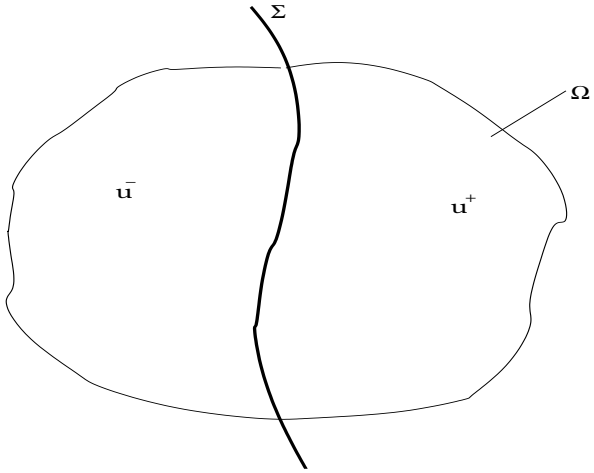

Figure 1.3: Jump discontinuity in the domain Ω divided by the surface Σ

from Ω^- to Ω^+. The following theorem gives a criterion which jump discontinuities are admissible across the surface Σ:

Theorem 1.6
Consider $u : \mathbb{R}^d \times [0,+\infty[\to \mathbb{R}$ as a C^1-function in the above sense. Then u is a weak solution of (1.10) if and only if the following conditions hold:

(i) u is a classical solution of (1.10) in the domains where it is C^1,

(ii) u satisfies the jump condition

$$[u^+ - u^-]n_t \;-\; \sum_{i=1}^{d} [f_i(u^+) - f_i(u^-)]n_{x_i} \;=\; 0 \tag{1.15}$$

along Σ with shock speed $\sigma := n_t$.

Proof [31] ∎

The Riemann problem

The Rankine-Hugoniot jump condition gives a criterion which solutions are admissible across a discontinuity. The question arises how this discontinuities propagate in space-time, i.e. by which solutions the constant values u^+ and u^- can be connected.

A physical realization of such a problem is the following: Assume an infinitely long tube filled with a gas in two different states separated by a membrane. At time $t = 0$ the membrane is removed. The question is how these states interact. Such a problem was first considered by Bernhard Riemann[2] [91], and therefore it is named Riemann problem.

The mathematical formulation of this problem is defined in the following way:

Definition 1.7
A conservation law (1.10) together with piecewise constant initial data separated by a single discontinuity Σ is known as the Riemann problem. *This consists of solving the Cauchy problem (1.10) with initial data*

$$u_0(\underline{x}) = \begin{cases} u^+, & \lim\limits_{\varepsilon\to 0} u((\underline{x},t) + \varepsilon\underline{n}), \\ u^-, & \lim\limits_{\varepsilon\to 0} u((\underline{x},t) - \varepsilon\underline{n}). \end{cases} \tag{1.16}$$

For a single space dimension the Riemann problem contains all solutions which are invariant under similarity transformations $(t,x) \mapsto (at, ax), a > 0$. More precisely we have to seek the solutions among the class of self-similar solutions of the form $u(t,x) = v(x/t)$. So the Riemann problem reduces to the ordinary differential equation

$$\begin{aligned} (f(v))'(\xi) &= \xi v'(\xi), \quad \xi \in \mathbb{R}^d, \\ v(-\infty) &= u^-, \\ v(+\infty) &= u^+, \end{aligned}$$

where the solution u is constant along the straight lines $\xi = x/t =$ constant and moreover has a simple structure. It consists of constant states separated by combinations of simple waves, either rarefaction or shock waves (see e.g. [101]):

Shock waves

A shock wave connecting the states $u^-(x,t)$ and $u^+(x,t)$ is a discontinuous solution of the Cauchy problem (1.10), (1.16) with

$$u(x,t) = \begin{cases} u^-, & x < \sigma t, \\ u^+, & x > \sigma t. \end{cases},$$

(see figure (1.2)). The shock speed σ is given by the Rankine-Hugoniot condition (1.15), i.e.

$$\sigma = \frac{f(u^+) - f(u^-)}{u^+ - u^-}$$

and $\underline{n} = (n_t, n_x)^T = (\sigma, id)^T$.

Example 1.8 (Shock waves [28])
We consider the canonical example for scalar conservation laws in one space dimension – the

[2]Bernhard Riemann, Göttingen (1826 – 1866)

Burgers' equation[3]*:*

$$\begin{aligned} \partial_t u + \partial_x \left(\frac{u^2}{2} \right) &= 0 \qquad in \quad (0,\infty) \times \mathbb{R} \\ u(0,x) &= u_0(x) \end{aligned} \tag{1.17}$$

which can be considered as the limit for $\varepsilon \searrow 0$ *of the parabolic equations*

$$\partial_t u + \partial_x \left(\frac{u^2}{2} \right) = \varepsilon \partial_x^2 u \tag{1.18}$$

called the viscous Burgers' equation.

We assume the initial conditions

$$u_0(x) = \begin{cases} 1 & x \le 0 \\ 1-x \quad if & 0 < x < 1 \\ 0 & x \ge 1 \end{cases} .$$

Since u *is constant along a characteristic line, i.e.* $u(t, x(t)) = u_0(x_p)$ *along the projected characteristic*

$$x(t) = u_0(x_p)t + x_p, \quad \forall x_p \in \mathbb{R},$$

one has

$$u(t,x) = \begin{cases} 1 & x \le t,\ 0 \le t < 1 \\ \dfrac{1-x}{1-t} \quad for & t \le x \le 1,\ 0 \le t < 1 \\ 0 & x \ge 1,\ 0 \le t < 1 \end{cases} .$$

The remarkable fact is, that this solution is only defined for $t \le 1$ *and breaks down for* $t > 1$ *since the characteristics then cross. We already have discussed such problems and know that a shock forms. Since we have* $u^- = 1, u^+ = 0$ *and* $f(u^-) = \frac{1}{2}(u^-)^2 = \frac{1}{2}, f(u^+) = 0$ *we are able to compute the Rankine-Hugoniot condition (1.15), i.e.*

$$\sigma = \frac{f(u^+) - f(u^-)}{u^+ - u^-} = \frac{0 - 1/2}{0 - 1} = \frac{1}{2}.$$

Therefore, we have for the curve Σ *parameterised by* $s(t)$*,*

$$\dot{s} = \sigma = \frac{1}{2},$$

and take $s(t) = \frac{1+t}{2}$*. The solution for the time* $t \ge 1$ *writes*

$$u(t,x) = \begin{cases} 1 & x < s(t) \\ \quad for & \\ 0 & s(t) < x \end{cases} .$$

[3]Due to a note by Dafermos [21] it was apparently Bateman [7] who first suggested that (1.17) and (1.18) should be employed as models for the system of conservation laws of inviscid and viscous gases. The commonly used name of Burgers [12] was attached to these equations by Hopf [53].

Rarefaction waves

A rarefaction wave or fan is a continuous solution of (1.10) connecting the states u^+ and u^- of the Riemann problem with $u^- < u^+$. Thus, the function v must satisfy the ordinary differential equation

$$\left[f'(v(\underline{\xi})) - \underline{\xi}\right] v'(\underline{\xi}) \quad = \quad 0.$$

Since we are only interested in solutions connecting constant states, we exclude those states corresponding to the solution $v'(\xi) = 0$ and obtain

$$f'(v(\xi)) \quad = \quad \xi.$$

If we assume the function f to be convex, i.e. $f'' > 0$ this defines a unique function $v(\xi)$. Thus, a rarefaction wave is defined by

Definition 1.9
Consider a Riemann problem (1.10) with initial data (1.16) and $u^- < u^+$. The solution connecting both states is called rarefaction wave *and is given by*

$$u(x,t) = \begin{cases} u^- & & & x/t \leq f'(u^-), \\ v(x/t) & \text{for} & f'(u^-) \leq x/t \leq f'(u^+), \\ u^+ & & f'(u^+) \leq x/t. \end{cases} \tag{1.19}$$

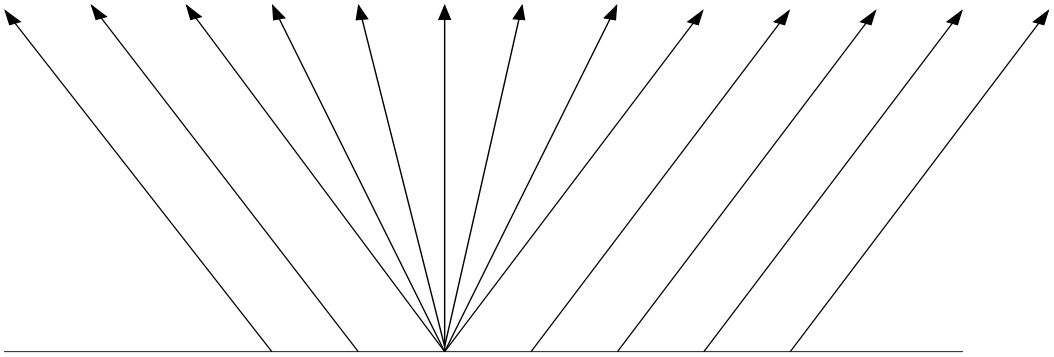

Figure 1.4: Characteristic lines for a rarefaction wave

Example 1.10 (Rarefaction wave and nonphysical shocks)
Consider the initial value problem (1.17) with data

$$u_0(x) \quad = \quad \begin{cases} 0 & & x < 0 \\ & \text{for} & \\ 1 & & x > 0 \end{cases}. \tag{1.20}$$

The method of constructing the characteristic lines does not lead to any ambiguity in defining u, but does fail to determine the solution inside the wedge $\{0 < x < t\}$. Consider the solution

$$u_1(t,x) \quad = \quad \begin{cases} 0 & & x < \dfrac{t}{2} \\ & \text{for} & \\ 1 & & x > \dfrac{t}{2} \end{cases}.$$

It is easy to see that the Rankine-Hugoniot condition holds and u is a weak solution of (1.17), (1.20). Nevertheless, this solution represents a nonphysical shock. *The reason why this is not an admissible solution will be given in the section concerning entropy solutions.*

However, one can create another solution given by

$$u_2(t,x) \quad = \quad \begin{cases} 0 & x < 0 \\ \dfrac{x}{t} & \text{for} \quad 0 < x < t \\ 1 & x > t \end{cases},$$

which is a rarefaction wave. This solution is also a continuous solution for (1.17), (1.20).

Contact discontinuity

A contact discontinuity occurs if the function f is affine on the interval limited by $u^-(x_0)$ and $u^+(x_0)$, i.e. $f'(u^-(x_0)) = f'(u^+(x_0)) = s$ which means that the characteristics run parallel to the discontinuity with speed s (see figure(1.5)).

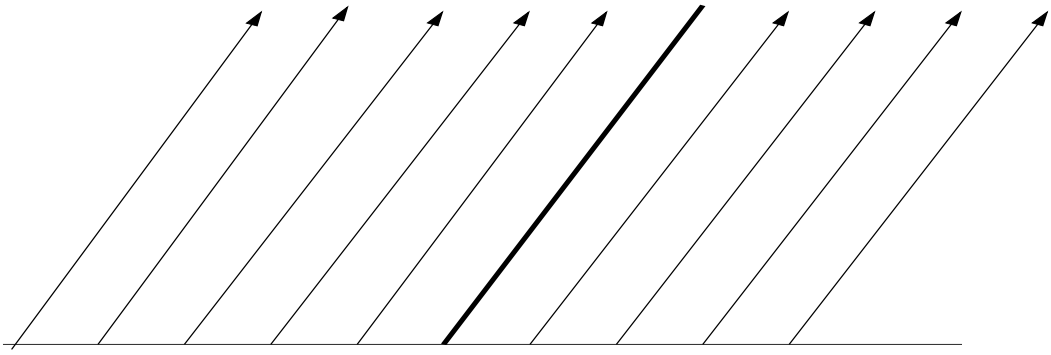

Figure 1.5: Contact discontinuity

The entropy condition

We have seen in the foregoing section, that a weak solution of the Cauchy problem (1.10) is not necessarily unique. Hence we need to find an additional criterion which enables us to find the physically relevant solution, at least when we consider physically relevant problems like conservation laws modelling physical principles. The use of entropy inequalities to characterise admissible solutions was proposed first by Kruzkov [62] and elaborated by Lax [66]. We start with an instructive explanation and proceed to a rigorous mathematical notion of entropy.

In the beginning of this chapter, we have seen how conservation laws arise as a mathematical abstraction from physical principles. So it seems justified to consider a scalar conservation law (1.10) as a simplification of a more complex case, e.g. as the one-sided limit for $\varepsilon \searrow 0$ of the dissipative case

$$\partial_t u^\varepsilon + \langle \nabla, \underline{f}(u^\varepsilon) \rangle \quad = \quad \varepsilon \Delta u^\varepsilon. \tag{1.21}$$

This model is called the diffusion-entropy approach. It describes the fact that conservation laws reveal a similar behaviour as physical systems controlled by entropy. If a classical solution of the differential equation (1.10) exists, it is invariant under parity-time-transformations (PT-transformations), i.e. changing signs of time and all space coordinates (parity). This reveals the reversibility of the system as long as we have smooth solutions.

For discontinuities, however, the PT-transformation is violated and the reversibility breaks down. This reveals the fact that if a characteristic wave propagates into a shock, it can not be traced back in space-time and determined where and when this had occurred. This information is lost inside the system. A mechanism similar to entropy production takes place, the formation of a shock wave is irreversible, cf. [4].

The physical analogue that one may have in mind is the movement of a particle along a streamline. Moving into a shock, the particle is slowed down, friction takes place and one has to look for dissipative mechanisms which transform the kinetic energy of the particle into heat. Loss of information and entropy increase takes place.

This model is revealed by the diffusion-entropy approach, where we consider an entropy, or more precisely an entropy-entropy flux pair $(\mathsf{u}, \underline{\mathsf{f}})$. Here, $\mathsf{u}(u)$ is an arbitrary strictly convex function and $\underline{\mathsf{f}}$ the entropy flux, connected with u and $\underline{f}$ through

$$\underline{\mathsf{f}}'(u) = \mathsf{u}'(u)\underline{f}'(u). \tag{1.22}$$

(1.22) is the compatibility relation[4]. If we multiply (1.21) by u', i.e.

$$\mathsf{u}'\partial_t u + \mathsf{u}'\langle \nabla, \underline{f}(u)\rangle \;=\; \mathsf{u}'\varepsilon\Delta u$$

we can rewrite this due to the compatibility relation mentioned as

$$\partial_t \mathsf{u} + \langle \nabla, \underline{\mathsf{f}}\rangle \;=\; \mathsf{u}'\varepsilon\Delta u.$$

Since $\mathsf{u}'\varepsilon\Delta u = \varepsilon[\nabla(\mathsf{u}'\nabla u) - \mathsf{u}''(\nabla u)^2]$, we have

$$\partial_t \mathsf{u} + \langle \nabla, \mathsf{f}\rangle \;=\; \varepsilon\nabla(\mathsf{u}'\nabla u) - \varepsilon\mathsf{u}''(\nabla u)^2.$$

With the convexity of u and $(\nabla u)^2 \geq 0$ we get

$$\partial_t \mathsf{u} + \langle \nabla, \mathsf{f}\rangle \;\leq\; \varepsilon\nabla(\mathsf{u}'\nabla u). \tag{1.23}$$

If we let ε tend to 0, we formally derive the inequality

$$\partial_t \mathsf{u} + \langle \nabla, \mathsf{f}\rangle \;\leq\; 0 \tag{1.24}$$

which is called the entropy inequality.

Existence and uniqueness of the limit function of (1.23) for $\varepsilon \searrow 0$ were given by Kruzkov [62] considering an entropy-entropy-flux pair of the form

$$\begin{aligned} \mathsf{u} &= |u-k|, \qquad k \in \mathbb{R}. \\ \mathsf{f} &= \operatorname{sign}(u-k)|f(u)-f(k)|. \end{aligned} \tag{1.25}$$

He showed that it is enough to consider this family of entropy-entropy-flux pairs to proof the entropy inequality for conservation laws. We cite the fundamental result for completeness:

[4]The reason for this notion will presented systematically in the case of multidimensional systems

Theorem 1.11
Assume that $u_0 \in L^\infty \cap BV(\mathbb{R}^d \times \mathbb{R} \to \mathbb{R})$ and the flux function $\underline{f} \in C^1(\mathbb{R}^d \to \mathbb{R})$ are Lipschitz continuous. Then u^ε converges almost everywhere in $[0,T]$ to a limit function $u \in L^\infty(\mathbb{R}^d \to \mathbb{R})$ for $\varepsilon \searrow 0$. This limit function is a unique entropy solution of (1.10).

Proof [31] ∎

1.3 Systems of conservation laws

In the following conservation laws for systems in several space dimensions are considered. To this use we follow the books of Majda [77] and Godlewski and Raviart [32]. The case of two space dimensions is treated by Lax [67] and Zheng [119].

Systems in several space dimensions

Consider some vector-valued functions $\underline{f}_i$, $1 \le i \le d$ of the form

$$\underline{f}_i : \begin{array}{ccc} \mathcal{S} & \to & \mathbb{R}^p, \\ \underline{u}(t,\underline{x}) & \mapsto & \underline{f}_i(\underline{u}(t,\underline{x})) \end{array} \qquad i = 1,\dots,d\,,$$

where $\underline{u} = (u_1,\dots,u_p)^T$ is a vector-valued function on $[0,+\infty[\times\mathbb{R}^d$ into $\mathcal{S}$, i.e.

$$\underline{u} : \begin{array}{ccc} \mathbb{R}_+ \times \mathbb{R}^d & \to & \mathcal{S}, \\ (t,\underline{x}) = (t,x_1,\dots,x_d) & \mapsto & u(t,\underline{x}). \end{array}$$

$\underline{u}$ is called as the vector of conserved quantities and the $\underline{f}_i(\underline{u})$ are known as the flux functions. Thus, the system of p conservation laws in d space variables reads as

$$\partial_t \underline{u} + \sum_{i=1}^{d} \partial_{x_i} \underline{f}_i(\underline{u}) \;=\; 0. \tag{1.26}$$

If the flux functions are continuously differentiable, we call (1.26) a system of conservation laws. The quasilinear form of (1.26) reads as

$$\partial_t \underline{u} + \sum_{i=1}^{d} \mathbf{A}_i \partial_{x_i} \underline{u} = 0. \tag{1.27}$$

where $\mathbf{A}_i$, $i = 1,\dots,d$ are the Jacobians of the flux functions $\underline{f}_i$, i.e.

$$\mathbf{A}_1(\underline{u}) = \nabla_{\underline{u}} \underline{f}_1(\underline{u}), \dots, \mathbf{A}_d(\underline{u}) = \nabla_{\underline{u}} \underline{f}_d(\underline{u}).$$

In general there is no problem to derive the quasilinear form (1.27) from (1.26), the converse is not true.

We consider the matrix

$$\mathbf{A}(\underline{u},\underline{n}) \quad = \quad \sum_{i=1}^{d} n_i \mathbf{A}_i(\underline{u}) \tag{1.28}$$

with $\underline{u} \in \Omega$ and $\underline{n} = (n_1, \ldots, n_d)^T \in \mathbb{R}^d$ a fixed unit vector. The eigenvalues of (1.28) are given by $\lambda_i(\underline{u},\underline{w})$. Then we have the following

Definition 1.12
If for the system (1.27) $\forall\, \underline{u} \in \mathcal{S}$, $\underline{n} \in \mathbb{R}^d$ with $\underline{n} \neq \underline{0}$, p real eigenvalues

$$\lambda_1(\underline{u},\underline{n}) \leq \ldots \leq \lambda_p(\underline{u},\underline{n})$$

and p linear independent right eigenvectors

$$\underline{r}_1(\underline{u},\underline{n}), \ldots, \underline{r}_p(\underline{u},\underline{n}),$$

of the matrix (1.28) exist, then the system is called hyperbolic.

If the eigenvalues of $\mathbf{A}(\underline{u},\underline{n})$, are real and distinct with respect to each other, i.e.

$$\lambda_1(\underline{u},\underline{n}) < \ldots < \lambda_p(\underline{u},\underline{n})$$

the system is called strictly hyperbolic.

Naturally, for distinct eigenvalues one has for every eigenvalue $\lambda_k(\underline{u},\underline{w})$ a left eigenvector $\underline{l}_k(\underline{u},\underline{w})$

$$\underline{l}_k(\underline{u},\underline{w})^T \mathbf{A}(\underline{u},\underline{w}) \quad = \quad \lambda_k(\underline{u},\underline{w}) \underline{l}_k(\underline{u},\underline{w})^T.$$

Thus, since the eigenvalues are distinct, $\{\underline{l}_k(\underline{u},\underline{n})\}_k$ form a dual basis of $\{\underline{r}_k(\underline{u},\underline{n})\}_k$, i.e.

$$\underline{l}_k(\underline{u},\underline{n})^T r_j(\underline{u},\underline{n}) \quad = \quad \langle l_k(\underline{u},\underline{n}), r_j(\underline{u},\underline{n})\rangle = 0, \quad k \neq j.$$

The quantities (eigenvalues, eigenvectors,...) associated with the index k are called k-th characteristic field. If we assume the flux-vectors $\underline{f}_i$, $i = 1, \ldots, d$ as C^2-functions they are C^1-functions of $\underline{u}$.

Travelling waves and hyperbolicity

In the scalar case, we already have seen in the example of the linear wave equation that a scalar conservation law models transport phenomena, i.e. the propagation of wave-like patterns.

Further on, we see that a similar solution for systems of conservation laws exists. For fixed vectors $\underline{u}_0 \in \mathcal{S}$ and $\underline{n} \in \mathbb{R}^d$ we consider solutions of the form $\underline{u}(t,\underline{x}) = \underline{u}_0 + \underline{v}$ and linearising around $\underline{u}_0 = \underline{0}$ we obtain the linearised system

$$\partial_t \underline{v} + \sum_{i=1}^{d} n_i \mathbf{A}(\underline{u}_0) \partial_{x_i} \underline{v} \quad = \quad 0. \tag{1.29}$$

This system possesses special forms of solutions called plane waves or travelling waves (see the classical textbook of John[5] [57] concerning travelling waves):

$$\underline{u}(\underline{x},t) \quad = \quad \sigma(\langle \underline{x},\underline{n}\rangle - \lambda_k(\underline{u}_0,\underline{n})t)\underline{r}_k(\underline{u}_0,\underline{n}), \tag{1.30}$$

where σ is an arbitrary scalar function of a scalar variable s. As one easily sees, this wave solutions involves only the k-th mode of the characteristic field propagating unidirectional, since $\underline{r}_k$ depends only on $\underline{u}_0$ and $\underline{n}$.

The question is how to obtain plane wave solutions for the nonlinear system. Thus, we consider solutions of the form

$$\underline{u}(t,\underline{x}) \quad = \quad \Gamma(\sigma(t,\langle \underline{x},\underline{n}\rangle)),$$

which define a curve in $\mathbb{R}^d$. We assume σ as a function of two scalar variables t and $\hat{x}$, i.e. $\sigma = \sigma(t,\hat{x})$. We have to choose $\Gamma(\sigma,\underline{n})$ such that it satisfies the nonlinear ordinary differential equation

$$\begin{aligned} \Gamma' \quad &= \quad \underline{r}_k(\Gamma(\sigma),\underline{n}), \\ \sigma(0) \quad &= \quad \underline{u}_0, \end{aligned} \tag{1.31}$$

i.e. Γ is an integral curve of $\underline{r}_k$. From this equation we derive (see Majda [77] for details)

$$\sigma(t,\hat{x}) \quad = \quad \sigma_0(\hat{x} - \lambda_k(\Gamma(\sigma,\underline{n}))t)$$

which is valid as long as σ remains smooth. This leads to the nonlinear planar wave solution of (1.26)

$$\underline{u}(t,\underline{x}) \quad = \quad \Gamma(\sigma_0(\langle \underline{x},\underline{n}\rangle - \lambda_k(\Gamma(\sigma))t),\underline{n}), \tag{1.32}$$

also known as k-simple wave. The question is: when the nonlinear solution behaves like the solution of the linear equation (1.30)? For the linear solution the important fact is that the shape of the waves is preserved for all times. This situation is true for the nonlinear solution (1.32) for a given $\underline{n}$ and for initial data $\sigma^- < \sigma_0(\underline{x}) < \sigma^+$ if and only if

$$\frac{d}{d\sigma}\lambda_k(\Gamma(\sigma,\underline{n})) \equiv 0.$$

According to (1.30) this is equivalent to the requirement

$$\langle \nabla_{\underline{u}}\lambda_k(\Gamma),\underline{r}_k\rangle\,\big|_{\Gamma(\sigma)} \equiv 0,$$

so that $\lambda_k(\Gamma(\sigma)) = \lambda_k(\underline{u}_0)$.

Definition 1.13
The k-th characteristic field *is said to be* linearly degenerated in the direction $\underline{n}$ *if*

$$\langle \nabla_{\underline{u}}\lambda_k(\underline{u},\underline{n}),\underline{r}_k(\underline{u},\underline{n})\rangle \quad = \quad 0$$

holds. We call the k-th wave field linearly degenerated if it is linearly degenerated for all $\underline{n}$ *with* $\|\underline{n}\| = 1$.

[5] Fritz John, Göttingen, Cambridge, New York (1910 – 1994)

The opposite situation takes place if the wave speed in (1.32) always changes with σ, i.e.

$$\frac{d}{d\sigma}\lambda_k(\Gamma(\sigma,\underline{n})) \neq 0, \quad \forall\sigma,\underline{u}_0.$$

This is equivalent to the condition

$$\frac{d}{d\sigma}\lambda_k(\Gamma(\sigma,\underline{n})) \;=\; \langle\nabla_{\underline{u}}\lambda_k(\underline{u},\underline{n}),\underline{r}_k(\underline{u},\underline{n})\rangle|_{\underline{u}=\Gamma(\sigma,\underline{n})} \;\neq\; 0$$

for (1.31). This leads to the

Definition 1.14
The k-th characteristic field *is said to be* genuinely nonlinear in the direction $\underline{n}$ *if*

$$\langle\nabla_{\underline{u}}\lambda_k(\underline{u},\underline{n}),\underline{r}_k(\underline{u},\underline{n})\rangle \;\neq\; 0, \qquad \forall\underline{u}\in\mathcal{S}$$

holds. We call the k-th wave field genuinely nonlinear *if it is genuinely nonlinear for all* $\underline{n}$, $\|\underline{n}\|=1$. *The genuinely nonlinear wave field is said to be* normalised, *if* $\underline{r}_k(\underline{u},\underline{n})$ *is scaled such that*

$$\langle\nabla_{\underline{u}}\lambda_k(\underline{u},\underline{n}),\underline{r}_k\rangle \;=\; 1.$$

Remark 1.15
Obviously, a k-th characteristic field is genuinely nonlinear, if $\nabla_{\underline{u}}\lambda_k(\underline{u},\underline{n})\neq 0$, *i.e.* $\lambda_k(\underline{u},\underline{n})$ *is a non-constant function of* $\underline{u}$, *and* $\nabla_{\underline{u}}\lambda_k(\underline{u},\underline{n})$ *is not orthogonal to* $\underline{r}_k$. *In the scalar case this is equivalent to* $\partial_u^2 f\neq 0$. *If* $\partial_u^2 f = 0$ *the characteristic curves* $dx/dt=\partial_u f$ *are all parallel. Thus, no intersection takes place and we have transport with constant speed without shock formation. This is the case for a linearly degenerated wave field.*

Characteristic curves

As in the scalar case, we try to express the conservation laws (1.26) as a directional derivative in space-time. Thus, we try to rewrite the conservation laws in p equations with $k=1,\ldots,p$:

$$\begin{aligned}
&\frac{\partial\underline{u}}{\partial t}+\sum_{i=1}^{d}\frac{\partial\underline{f}_i(\underline{u})}{\partial x_i} \;=\; 0\\
\Leftrightarrow\quad &\frac{\partial u_k}{\partial t}+\sum_{i=1}^{d}\left(\sum_{j=1}^{p}\frac{\partial f_{i,k}(\underline{u})}{\partial u_j}\frac{\partial u_j}{\partial x_i}\right) \;=\; \frac{\partial u_k}{\partial t}+\sum_{j=1}^{p}\left(\sum_{i=1}^{d}\frac{\partial f_{i,k}(\underline{u})}{\partial u_j}\frac{\partial u_j}{\partial x_i}\right)\\
&\qquad\qquad = \frac{\partial u_k}{\partial t}+\sum_{j=1}^{p}\left(\frac{\partial f_{1,k}(\underline{u})}{\partial u_j},\ldots,\frac{\partial f_{d,k}(\underline{u})}{\partial u_j}\right)^T\nabla u_j\\
&\qquad\qquad = \sum_{j=1}^{p}\left(\frac{\partial u_k}{\partial t}\delta_{kj}+\frac{\partial f_{1,k}(\underline{u})}{\partial u_j},\ldots,\frac{\partial f_{d,k}(\underline{u})}{\partial u_j}\right)^T\nabla u_j\\
&\qquad\qquad = \sum_{j=1}^{p}\left(\delta_{kj},\frac{\partial f_{1,k}(\underline{u})}{\partial u_j},\ldots,\frac{\partial f_{d,k}(\underline{u})}{\partial u_j}\right)^T\nabla^{\mathsf{t}}u_j \;=\; 0.
\end{aligned}$$

For ease of notation we have written the inner product $\langle \underline{x}, \underline{y} \rangle$ for two arbitrary vectors $\underline{x}, \underline{y} \in \mathbb{R}^{d+1}$ as $\underline{x}^T \underline{y}$. The outer product used later in this context reads as $\underline{x}\,\underline{y}^T$. The symbol δ_{kj} – the Kronecker[6]-delta – is defined as

$$\delta_{kj} \;:=\; \begin{cases} 0 & k \neq j, \\ & \text{if} \\ 1 & k = j. \end{cases}$$

We obtain a system of quasilinear partial differential equations of the form

$$\frac{\partial u_k}{\partial t} + \sum_{i=1}^{d} \frac{\partial \underline{f}_{i,k}(\underline{u})}{\partial x_i} \;=\; \underbrace{\sum_{j=1}^{p} [\underline{a}_{kj}(\underline{u})]^T \nabla^{\mathsf{t}} u_j}_{:=\pi_k} \;=\; 0, \quad k = 1, \ldots, p, \tag{1.33}$$

with

$$[\underline{a}^{\mathsf{t}}_{kj}(\underline{u})]^T \;=\; \left(\delta_{kj}, \frac{\partial f_{1,k}(\underline{u})}{\partial u_j}, \ldots, \frac{\partial f_{d,k}(\underline{u})}{\partial u_j} \right).$$

Thus, we get p equations each representing linear combinations of derivatives of the conserved quantity u_k into the direction of $\underline{a}^{\mathsf{t}}_{kj}(\underline{u})$.

As in the scalar case, we look for distinguished directions for derivatives. Müller [83] considers three different solution approaches to this question, where every Ansatz leads to the same result. We only consider the third approach which is similar to the proceeding in the scalar case:

– Combine the differential equations (1.33) in such manner that derivation is only allowed inside the hyperplane Σ, not normal to it. (Reduction of the derivative directions).

Consider the vector $\underline{n}^{\mathsf{t}}$ as the vector normal to the hyperplane Σ^{t}. The differential equations π_k, $k = 1, \ldots, p$ should be combined in such a way that one direction of derivation disappears, i.e. we seek a vector $\underline{\eta} \in \mathbb{R}^p \backslash \{\underline{0}\}$ meeting

$$\underline{\eta}\Pi \;=\; \sum_{k=1}^{p} \eta_k \pi_k \;=\; \sum_{k=1}^{p} [\underline{b}^{\mathsf{t}}_k]^T \nabla^{\mathsf{t}} u_k \;=\; 0, \tag{1.34}$$

with

$$[\underline{b}^{\mathsf{t}}_k]^T \;=\; \sum_{j=1}^{p} \eta_k [\underline{a}^{\mathsf{t}}_{kj}]^T, \quad k = 1, \ldots, p, \tag{1.35}$$

which are nothing more than the directions of derivative we are looking for. Since all these directions should lie only in Σ^{t}, they have to be perpendicular to $\underline{n}^{\mathsf{t}}$ which leads to

$$\det\left(\underline{n}^{\mathsf{t}} [\underline{a}^{\mathsf{t}}_{kj}]^T \right) \;=\; 0.$$

[6] Leopold Kronecker, Berlin $(1823 - 1891)$

From this condition we can compute the normal vector $\underline{n}^{\mathsf{t}}$. In addition, it assures that a nontrivial linear combination exists, i.e. $\underline{\eta} \neq \underline{0}$. Solving the linear system of equations

$$\underline{\eta}^T \left(\underline{n}^{\mathsf{t}} [\underline{a}_{kj}^{\mathsf{t}}]^T \right) \quad = \quad \underline{0}$$

yields $\underline{\eta}$. As in the scalar case, we diminish the directions of derivation by one direction. There we could reduce the partial differential equation to an ordinary one which is solved by the characteristic curve. In the case of a system, the solution is given by characteristic hyperplanes. For the most prominent example of a hyperbolic system, the Euler[7]-equations, two characteristic hyperplanes exist:

- trajectory 1-dim hyperplane
- sound propagation d-dim hyperplane

The Riemann problem and Riemann invariants

Now we turn to the Riemann problem for systems. As in the scalar case this denotes an initial value problem with two constant states separated by a hyperplane:

Definition 1.16
A hyperbolic system of the form (1.26), i.e.

$$\partial_t \underline{u} + \sum_{i=1}^{d} \partial_{x_i} \underline{f}_i(\underline{u}) = 0.$$

with initial data

$$\underline{u}_0(\underline{x}) \; = \; \underline{u}(t_0, \underline{x}) \quad = \quad \begin{cases} \underline{u}^-, & \langle \underline{x}, \underline{n} \rangle \leq \langle \underline{x}_0, \underline{n} \rangle, \\ \underline{u}^+, & \langle \underline{x}, \underline{n} \rangle > \langle \underline{x}_0, \underline{n} \rangle, \end{cases} \tag{1.36}$$

and fixed unit vector $\underline{n} \in \mathbb{R}^d$ is called Riemann problem for systems.

Lax [68] showed that for small discontinuities in $\underline{u}_0$ a solution always exists. For the scalar case this is independent of the size of the jump. An explicit formula for the weak solution is available. In the case of a system the situation is more involved.

Lax solved the problem by defining $p-1$ intermediate states $\underline{u}_{s_1}, \ldots, \underline{u}_{s_{p-1}} \in \mathcal{S}$ and a path $\Gamma_{\mathcal{S}} : \mathbb{R} \to \mathcal{S}$ connecting $\underline{u}^- =: \underline{u}_{s_0}, \underline{u}_{s_1}, \ldots, \underline{u}_{s_{p-1}}, \underline{u}_{s_p} := \underline{u}^+$. This is done in such a way that two subsequent states $\underline{u}_{s_{k-1}}, \underline{u}_{s_k}$ are joined by a subpath $\Gamma_{\mathcal{S}_k}$ representing either a rarefaction wave, a shock or a contact discontinuity. The linear independent eigenvectors $\underline{r}_1, \ldots, \underline{r}_p$ are tangent to these subpaths. To compute the intermediate states $\underline{u}_{s_i}, i = 1, \ldots, p-1$ functions are introduced. The important fact is, that these functions are constant along the respective paths. These are the Riemann invariants:

Definition 1.17
A continuously differentiable function $\mathsf{R} : \Omega \to \mathbb{R}$ is called k-th Riemann invariant, *if*

$$\langle \underline{r}_k(\underline{u}), \nabla_{\underline{u}} \mathsf{R}(\underline{u}) \rangle \quad = \quad 0.$$

[7] Leonhard Euler, Basel, Berlin, St. Petersburg (1707 – 1783)

Since we are looking for Riemann invariants along the paths connecting the above mentioned intermediate states we have

Lemma 1.18
A k-th Riemann invariant R *is constant along a path* $\Gamma_k : s \in \mathbb{R} \to \Gamma_k(s)$ *with*

$$\frac{d}{ds}\Gamma_k \quad = \quad \underline{r}_k \circ \Gamma_k, \qquad \Gamma_k(0) = \underline{\hat{u}} \tag{1.37}$$

for arbitrary $\underline{\hat{u}} \in \mathcal{S}$.

This is a path of the same form as (1.31) so that the existence and uniqueness for the open interval $I \in \mathbb{R}$ containing the point $s = 0$ is assured. For a k-th Riemann invariant we get

$$\frac{d}{ds}(\mathsf{R} \circ \Gamma_k) \quad = \quad \langle \nabla_{\underline{u}} \mathsf{R}, \underline{r}_k \rangle \circ \Gamma_k = 0 \tag{1.38}$$

and consequently

$$\mathsf{R}(\Gamma_k(s)) = \mathsf{R}(\Gamma_k(0)) = \mathsf{R}(\underline{\hat{u}}).$$

Here, we already see that for a linearly degenerated field, the eigenvalue λ_k is a k-th Riemann invariant because we have

$$\langle \nabla_{\underline{u}} \lambda_k(\underline{u}, \underline{n}), \underline{r}_k(\underline{u}, \underline{n}) \rangle \quad = \quad 0.$$

Remark 1.19
For a fixed $k \in \{1, \ldots, p\}$ *there are* $p-1$ *k-th Riemann invariants since in* $\mathbb{R}^p$ *there are* $p-1$ *directions orthogonal to the tangent vector* $\underline{r}_k$. *Thus, we derive* $p(p-1)$ *Riemann invariants to compute the* $p(p-1)$ *unknowns of the* $p-1$ *intermediate states mentioned above.*

Thus, we sum the remarks above in the following Theorem which is proved in [32]:

Theorem 1.20
Assume that for all $k \in 1, \ldots, p$ *the k-th characteristic field is either genuinely nonlinear or linearly degenerated. Then for all* $\underline{u}^- \in \mathcal{S}$ *there exists a neighbourhood* ϑ *of* $\underline{u}^-$ *in* $\mathcal{S}$ *with the following property: If* $\underline{u}^+$ *belongs to* ϑ, *the Riemann-problem (1.26)(1.36) has a weak solution that consists of at most (p+1) constant states separated by rarefaction waves, admissible shock waves, or contact discontinuities. Moreover, a weak solution of this kind is unique.*

Rarefaction waves

As in the scalar case, a rarefaction wave is a centred wave connecting the constant states $\underline{u}^-$ and $\underline{u}^+$ of the Riemann-problem (1.36).

Thus, we start from the assumption that the k-th wave field may be genuinely nonlinear and normalised for a path $\Gamma_k(s)$, i.e.

$$\frac{d}{ds}\lambda_k(\underline{u})\Big|_{\Gamma_k(s)} \quad = \quad \langle \nabla_{\underline{u}} \lambda_k(\Gamma_k), \underline{r}_k(\Gamma_k) \rangle \; = \; 1$$

and we get

$$\lambda_k(\Gamma_k(s)) = \lambda_k(\Gamma_k)(0) + s = \lambda_k(\underline{\hat{u}}) + s.$$

We put $\underline{\hat{u}} := \underline{u}^-$ in (1.37), set

$$\xi := \langle \tfrac{\underline{x}-\underline{x}_0}{t-t_0}, \underline{n}\rangle - \lambda_k(\underline{u}_{k-1}, \underline{n}),$$

and define for an arbitrary state $\underline{u}^+ \in \Gamma_k \in (I \cap \mathbb{R}_{>0})$

$$\underline{u}(t,\underline{x}) \;:=\; \begin{cases} \underline{u}^-, & \xi < 0, \\ \Gamma_k(\xi), & 0 \le \xi \le \lambda_k(\underline{u}_k, \underline{n}) - \lambda(\underline{u}_{k-1}, \underline{n}) \\ \underline{u}^+, & \xi > \lambda_k(\underline{u}_k, \underline{n}) - \lambda(\underline{u}_{k-1}, \underline{n}). \end{cases} \tag{1.39}$$

Definition 1.21
A self-similar weak solution (1.39) of (1.26) is called a k-th centred simple wave *or* k-th rarefaction wave *connecting the states* $\underline{u}^-, \underline{u}^+ \in \mathcal{S}$.

That (1.39) is a solution of the system (1.26) can be seen in the following way. If we differentiate the middle line of (1.39) with respect to time we have

$$\left.\frac{\partial}{\partial t}\right|_{t,\underline{x}} u \;=\; -\langle \tfrac{\underline{x}-\underline{x}_0}{(t-t_0)^2}, \underline{n}\rangle \; \tfrac{d}{ds}\big|_\xi \, \Gamma_k.$$

Using

$$\nabla|_{t,\underline{x}} \underline{u} \;=\; \frac{\underline{n}}{t-t_0} \left.\frac{d}{ds}\right|_\xi \Gamma_k,$$

one gets

$$\begin{aligned} \mathbf{A}(\underline{n})|_{\Gamma_k(s)} \nabla|_{t,\underline{x}} \underline{u} &= \mathbf{A}(\underline{n})|_{\Gamma_k(s)} \frac{\underline{n}}{t-t_0} \left.\frac{d}{ds}\right|_\xi \Gamma_k \\ &= \frac{\underline{n}}{t-t_0} A(\underline{n})|_{\Gamma_k(s)} \underline{r}_k(\Gamma_k) \\ &= \frac{\underline{n}}{t-t_0} \lambda_k(\Gamma_k(s)) \underline{r}_k(\Gamma_k) \\ &= \frac{\underline{n}}{t-t_0} \lambda_k(\Gamma_k(s)) \left.\frac{d}{ds}\right|_\xi \Gamma_k \end{aligned}$$

and with the above definitions

$$= \frac{\underline{n}}{t-t_0} \left(\lambda_k(\underline{u}_{k-1}) + \langle \tfrac{\underline{x}-\underline{x}_0}{t-t_0}, \underline{n}\rangle - \lambda_k(\underline{u}_{k-1}) \right) \left.\frac{d}{ds}\right|_\xi \Gamma_k,$$

which yields

$$\sum_{i=1}^{d} \nabla_{\underline{u}} \underline{f}_i|_{t,\underline{x}} \partial_{x_i} \underline{u} \;=\; \langle \tfrac{\underline{x}-\underline{x}_0}{(t-t_0)^2}, \underline{n}\rangle \; \tfrac{d}{ds}\big|_\xi \, \Gamma_k = -\partial_t u|_{t,\underline{x}}.$$

Since the $\underline{u}_{k-1}$ and $\underline{u}_k$ are constant states, which naturally satisfy (1.27), this proves that (1.39) is a solution of (1.26).

Concerning the behaviour of the Riemann invariants for centred waves we present the following

Theorem 1.22
On a k-th rarefaction wave, all k-th Riemann invariants are constant

Proof Let $\underline{u}$ be a k-th rarefaction wave of the form (1.39), and let R be a k-th Riemann-invariant. The function

$$\mathsf{R}(\underline{u}) \quad : \quad (t, \underline{x}) \to \mathsf{R}(\underline{u}(t, \underline{x})), \quad t > 0$$

is continuous. For $\xi < 0$ and $\xi > \lambda_k(\underline{u}^+) - \lambda_k(\underline{u}^-)$, $\mathsf{R}(\underline{u})$ is constant. For $0 \leq \xi \leq \lambda_k(\underline{u}^+) - \lambda_k(\underline{u}^-)$, $\underline{u} = \Gamma_k(\xi)$ is an integral curve of $\underline{r}_k$ which proves the result. ∎

Discontinuity waves

Since we have defined rarefaction waves connecting two arbitrary states $\underline{u}^-, \underline{u}^+ \in \mathcal{S}$ continuously, we are now looking for discontinuity waves – in detail shock waves and contact discontinuities – that connect both states by a discontinuous solution of (1.26).

Recall that along a discontinuity Σ the Rankine-Hugoniot jump condition holds, i.e.

$$\langle [\underline{u}^+ - \underline{u}^-], \underline{n} \rangle \sigma \quad = \quad \sum_{i=1}^{d} [\underline{f}_i(\underline{u}^+) - \underline{f}_i(\underline{u}^-)] n_j \tag{1.40}$$

where $\underline{n} \in \mathbb{R}^d$ is the vector normal to the hypersurface Σ. Similar to the scalar case $\sigma \in \mathbb{R}$ plays the role of a shock speed, which will be defined in a minute.

In our discussion above, we mentioned that a given state $\underline{u}^-$ can be connected with states $\underline{u} \in \mathcal{S}$ by p smooth curves $\Gamma_k(\underline{u}^-), 1 \leq k \leq p$. Thus, we give the following

Definition 1.23
The Rankine-Hugoniot set *of states $\underline{u}^-$ is the set of all $\underline{u} \in \mathcal{S}$ such that there exists a $\sigma(\underline{u}^-, \underline{u}) \in \mathbb{R}$ satisfying the Rankine-Hugoniot jump condition (1.40).*

Thus, if we are looking for all states $\underline{u}^+$ which can be connected to $\underline{u}^-$ by a discontinuity wave, these states are included in this set. The existence of these paths is given by the following theorem whose proof can be found in [32]:

Theorem 1.24
Let $\underline{u}^-$ be in $\mathcal{S}$. The Rankine-Hugoniot set *of $\underline{u}^-$ is locally made of p smooth curves $\Gamma_k(\underline{u}^-), 1 \leq k \leq p$. Furthermore, there exists a parameterisation of $\Gamma_k(\underline{u}^-) : s \to \Gamma_k(s)$ defined for $|s| \leq s_1, s_1$ small enough, such that*

$$\Gamma(s) = \underline{u}^- + s\underline{r}_k(\underline{u}^-) + \frac{s^2}{2}\langle \nabla r_k(\underline{u}^-), \underline{r}_k(\underline{u}^-) \rangle + O(s^3) \tag{1.41}$$

and

$$\sigma(\underline{u}^-, \Gamma_k(s)) = \lambda_k(\underline{u}^-) + \frac{s}{2}\langle \nabla r_k(\underline{u}^-), \underline{r}_k(\underline{u}^-) \rangle + O(s^2). \tag{1.42}$$

The equations (1.41),(1.42) have some interesting consequences. We observe the behaviour of a k-th Riemann invariant along such a path Γ_k. Differentiating relation (1.38) one gets

$$\langle \nabla, \nabla \mathsf{R}(\underline{u})\rangle\langle \underline{r}_k(\underline{u}), \underline{v}\rangle + \langle \nabla, \underline{r}_k(\underline{u})\rangle\langle \nabla \mathsf{R}(\underline{u}), \underline{v}\rangle \;=\; 0.$$

By using (1.41), we obtain

$$\begin{aligned}\mathsf{R}(\Gamma_k(s)) &= \mathsf{R}\left(\underline{u}^- + s\underline{r}_k(\underline{u}^-) + \frac{s^2}{2}\langle \nabla r_k(\underline{u}^-), \underline{r}_k(\underline{u}^-)\rangle + O(s^3)\right)\\ &= \mathsf{R}(\underline{u}^-) + s\nabla\mathsf{R}(\underline{u}^-) + \frac{s^2}{2}\left[\langle \nabla, \nabla\mathsf{R}(\underline{u}^-)\rangle\ \langle \underline{r}_k(\underline{u}^-), \underline{r}_k(\underline{u}^-)\rangle\right.\\ &\qquad \left. + \langle \nabla, \underline{r}_k(\underline{u}^-)\rangle\langle \nabla\mathsf{R}(\underline{u}^-), \underline{r}_k(\underline{u}^-)\rangle\right] + O(s^3).\end{aligned}$$

Thus, this proves

Corollary 1.25
For a k-th Riemann invariant along a path Γ_k

$$\mathsf{R}(\Gamma_k(s)) \;=\; \mathsf{R}(\underline{u}^-) + O(s^3)$$

holds.

If we consider the case of a genuinely nonlinear k-th characteristic field, we call the path $\Gamma_k(\hat{\underline{u}})$ a k-shock curve and derive the following

Definition 1.26
If $\underline{u}^+$ belongs to the k-shock curve $\Gamma_k(\underline{u}^-)$, or equivalently if $\underline{u}^-$ belongs to the k-shock curve $\Gamma_k(\underline{u}^+)$, a weak solution of (1.26) of the form (1.36), (1.40) is called a k-shock wave *and has shock speed of the form*

$$\sigma(\underline{u}^-, \underline{u}^+) \;=\; \frac{1}{2}\left(\lambda_k(\underline{u}^-) + \lambda_k(\underline{u}^+)\right) + O(s^2). \tag{1.43}$$

Remark 1.27
The k-shock speed (1.43) is a direct consequence of applying (1.41) to $\lambda_k(\Gamma(s))$ and combining with (1.42) by setting $\hat{\underline{u}} = \underline{u}^-$ and $\Gamma_k(s) = \underline{u}^+$.

Having examined discontinuity waves in the genuinely nonlinear case as a k-shock wave, we now turn to the linearly degenerated case and obtain

Theorem 1.28
If the k-th characteristic field is linearly degenerated, the curve given by Lemma 1.18 is an integral curve for the vector field $\underline{r}_k$ and we obtain for the propagation of the discontinuity

$$\sigma(\hat{\underline{u}}, \Gamma_k(s)) \;=\; \lambda_k(\Gamma_k(s)) = \lambda_k(\hat{\underline{u}}).$$

For a k-th Riemann-invariant we have

$$\mathsf{R}(\Gamma_k(s)) = \mathsf{R}(\hat{\underline{u}}).$$

This result is also proved in [32] which leads to the following

Definition 1.29
Consider the k-th characteristic field as linearly degenerated and $\underline{u}^+ \in \Gamma_k(\underline{u}^-)$ *(or equivalently* $\underline{u}^- \in \Gamma_k(\underline{u}^+)$*). Then a weak solution of (1.26) of the form (1.36), (1.40), with*

$$\nu = \lambda_k(\underline{u}^-) = \lambda(\underline{u}^+) = \overline{\lambda},$$

i.e.

$$\underline{u}(t,\underline{x}) = \begin{cases} \underline{u}^-, & \langle \underline{x}, \underline{n} \rangle < \overline{\lambda} t, \\ \underline{u}^+, & \langle \underline{x}, \underline{n} \rangle > \overline{\lambda} t, \end{cases} \tag{1.44}$$

is called a k-contact discontinuity.

Godlewski and Raviart note that the solution (1.44) is the limit of a k-simple (noncentred) wave (1.32).

Symmetric systems

We already have stated some properties for systems of conservation laws in several space dimensions. But in fact, for such general systems very little is known, unless they are symmetrisable.

Definition 1.30
A system of conservation laws of the form (1.27) is called symmetrisable, if there exists a positive definite symmetric matrix

$$\mathbf{A}_0(\underline{u}) \in \mathbb{R}^{p \times p}, \quad \forall \underline{u} \in \mathcal{S},$$

smoothly varying with $\underline{u}$ *such that the matrix*

$$\mathbf{A}_0(\underline{u})\mathbf{A}_j(\underline{u})$$

is symmetric.

Friedrichs made the observation that all equations of classical physics, which can be cast in the form (1.27), are symmetrisable in the following sense:

Theorem 1.31
For all $\underline{u} \in \mathcal{S}$ *there is a symmetric matrix* $\mathbf{A}_0(\underline{u})$*, varying smoothly with* $\underline{u}$ *such that*

i) $c\mathbf{I} \leq \mathbf{A}_0(\underline{u}) \leq c^{-1}\mathbf{I}$*, with a constant* c *uniform for* $\underline{u} \in G_1$ *and any* G_1 *with* $\overline{G_1} \subset \mathcal{S}$*,*

ii) $\mathbf{A}_0(\underline{u})\mathbf{A}_j(\underline{u}) = \tilde{\mathbf{A}}_j(\underline{u})$ *with* $\tilde{\mathbf{A}}_\mathbf{j} = \tilde{\mathbf{A}}_\mathbf{j}^T, \quad j = 1, \ldots, d.$

Definition 1.32
Systems with the above property are called symmetrisable in the sense of Friedrichs.

The fact that conservation laws are symmetrisable is quite remarkable and closely related to find physically relevant weak solutions called entropy solutions. The notion of entropy will be treated in the following in depth. Here, we state some important relations between entropy functions and symmetric systems.

Theorem 1.33
Let $\mathsf{u} : \Omega \to \mathbb{R}$ *be a strictly convex function. A necessary and sufficient condition for* u *to be an entropy for the system (1.26) is that the matrices*

$$\nabla^2_{\underline{u}} \mathsf{u}(\underline{u}) \nabla_{\underline{u}} \underline{f}_i(\underline{u}), \in \mathbb{R}^{p \times p}, \quad 1 \leq i \leq d,$$

are symmetric.

This result goes back to Lax and Friedrichs [30] and is called Lax-Friedrichs symmetrisation. Due to the fact that they use the quasilinear form (1.27), this symmetrisation only conserves classical solutions, but not weak ones.

Mock [82] showed that one can symmetrise the system by introducing new dependent variables $\underline{v}$, i.e. $\underline{u} = \underline{u}(\underline{v})$. The system (1.26) then becomes

$$\nabla_{\underline{v}} \underline{u} \partial_t \underline{v} + \sum_{i=1}^{d} \nabla_{\underline{u}} \underline{f}_i \nabla_{\underline{v}} \underline{u} \partial_{x_i} \underline{v} \quad = \quad \underline{0},$$

and is symmetrised if $\nabla_{\underline{v}} \underline{u}(\underline{v})$ is a symmetric positive definite matrix and the matrices $\nabla_{\underline{u}} \underline{f}_i \nabla_{\underline{v}} \underline{u} \in \mathbb{R}^{p \times p}, 1 \leq i \leq d$ are symmetric.

Lemma 1.34
Consider $\mathsf{u}(\underline{u})$ *as a strictly convex function. Then the change of variables to* entropy variables *is defined by*

$$\underline{v}^T \quad = \quad \nabla_{\underline{u}} \mathsf{u}(\underline{u}),$$

and the compatibility relation is given by

$$\underline{v}^T \nabla_{\underline{v}} \underline{f}_i(\underline{u}(\underline{v})) \quad = \quad \nabla_{\underline{v}} \mathsf{f}_i(\underline{u}(\underline{v})). \tag{1.45}$$

Proof We assume u as strictly convex function, so the mapping $\underline{u} \to \nabla_{\underline{u}} \mathsf{u}(\underline{v})$ is one-to-one. Hence, the mapping is invertible and one ends up with the entropy variables

$$\underline{v}^T \quad = \quad \nabla_{\underline{u}} \mathsf{u}(\underline{u}).$$

Stated this, the compatibility relation (1.45) can be written as

$$\nabla_{\underline{u}} \mathsf{f}_i(\underline{u}) \quad = \quad \underline{v}(\underline{u})^T \nabla_{\underline{u}} \underline{f}_i(\underline{u}). \tag{1.46}$$

On the other hand we have

$$\begin{aligned} \nabla_{\underline{v}} \mathsf{f}_i(\underline{u}(\underline{v})) &= \nabla_{\underline{u}} \mathsf{f}_i(\underline{u}(\underline{v}))^T \nabla_{\underline{v}} \underline{u}, \\ \nabla_{\underline{v}} f_i(\underline{u}(\underline{v})) &= \nabla_{\underline{u}} f_i(\underline{u}(\underline{v}))^T \nabla_{\underline{v}} \underline{u}. \end{aligned}$$

Multiplying (1.46) with $\nabla_{\underline{v}} \underline{u}$ yields the proposed form. ∎

Theorem 1.35
A necessary and sufficient condition for the system (1.26) to possess a strictly convex entropy u *is that there exists a change of dependent variables that symmetrises (1.26).*

Proof We already have proven that a strictly convex entropy $\mathsf{u}(\underline{u})$ defines the change of variables as $\underline{v}^T = \nabla_{\underline{u}}\mathsf{u}(\underline{u})$.

Next we define the conjugate functions $\mathsf{u}^\star$ of u and $\mathsf{f}_i^\star$ of f_i, $1 \le i \le d$ by

$$\begin{aligned} \mathsf{u}^\star(\underline{v}) &:= \underline{v}^T\underline{u}(\underline{v}) - \mathsf{u}(\underline{u}(\underline{v})), \\ \mathsf{f}_i^\star(\underline{v}) &:= \underline{v}^T\underline{f}_i(\underline{v}) - \mathsf{f}_i(\underline{u}(\underline{v})). \end{aligned}$$

Differentiating $\mathsf{u}^\star$ and $\mathsf{f}_i^\star$ we obtain

$$\begin{aligned} \nabla_{\underline{v}}\mathsf{u}^\star(\underline{v}) &= \underline{u}(\underline{v})^T - \underline{v}^T\nabla_{\underline{v}}\underline{u}(\underline{v}) - \nabla_{\underline{u}}\mathsf{u}(\underline{u}(\underline{v}))\nabla_{\underline{v}}\underline{u}(\underline{v}) \\ &= \underline{u}(\underline{v})^T \end{aligned}$$

and

$$\begin{aligned} \nabla_{\underline{v}}\mathsf{f}_i^\star(\underline{v}) &= \underline{f}_i(\underline{u}(\underline{v}))^T + \underline{v}^T\nabla_{\underline{v}}\underline{f}_i(\underline{u}(\underline{v}))\nabla_{\underline{v}}\underline{u}(\underline{v}) - \nabla_{\underline{u}}\mathsf{f}_i(\underline{u}(\underline{v}))\nabla_{\underline{v}}\underline{u}(\underline{v}) \\ &= \underline{f}_i(\underline{u}(\underline{v})). \end{aligned}$$

Moreover, the matrices

$$\begin{aligned} \nabla_{\underline{v}}\underline{u}(\underline{v}) &= \nabla^2_{\underline{v}}\mathsf{u}^\star(\underline{v}), \\ \nabla_{\underline{v}}\underline{f}_i(\underline{u}(\underline{v}))\nabla_{\underline{v}}\underline{u}(\underline{v}) &= \nabla^2_{\underline{v}}\mathsf{f}_i^\star(\underline{v}), \quad 1 \le i \le d, \end{aligned} \tag{1.47}$$

are symmetric and

$$\nabla_{\underline{v}}\underline{u}(\underline{v}) = \left(\nabla^2_{\underline{v}}\mathsf{u}^\star(\underline{v})\right)^{-1}$$

is positive definite. This fact proves that the change of variables $\underline{v}^T = \nabla_{\underline{u}}\mathsf{u}(\underline{u})$ symmetrises (1.26).

Conversely, the symmetry of the matrices (1.47) implies the existence of $d+1$ functions $q(\underline{v}), p_i(\underline{v})$, $1 \le i \le d$, such that

$$\begin{aligned} \nabla q(\underline{v}) &= \underline{u}(\underline{v})^T, \\ \nabla p_i(\underline{v}) &= \underline{f}_i(\underline{u}(\underline{v})), \quad 1 \le i \le d. \end{aligned}$$

The positive-definiteness of $\nabla_{\underline{v}}\underline{u}(\underline{v})$ is equivalent to the strict convexity of the function $q(\underline{v})$. This implies that the mapping $\underline{v} \to \nabla_{\underline{v}}q(\underline{v})$ is one-to-one. So, $\underline{v}$ is a function of $\underline{u}$ and we write

$$\begin{aligned} \mathsf{u}(\underline{u}) &= \langle \underline{v}(\underline{u}), \underline{u}\rangle - q(\underline{v}(\underline{u})), \\ \mathsf{f}_i(\underline{u}) &= \langle \underline{v}(\underline{u}), \nabla\underline{f}_i(\underline{u})\rangle - p_i(\underline{v}(\underline{u})),\ 1 \le i \le d. \end{aligned}$$

Differentiating with respect to $\underline{u}$ gives

$$\nabla\mathsf{u}(\underline{u}) = \underline{v}(\underline{u}) + \langle \nabla, \underline{v}(\underline{u})\rangle[\underline{u} - \nabla q(\underline{v}(\underline{u}))] = \underline{v}(\underline{u})$$

and

$$\nabla\mathsf{f}_i(\underline{u}) = \langle \nabla, \underline{f}_i(\underline{u})\rangle + \langle \nabla, \underline{v}(\underline{u})\rangle[\underline{f}_i(\underline{u}) - \nabla p_i(\underline{u})] = \langle \nabla, \underline{f}_i(\underline{u})\rangle\underline{v}(\underline{u}).$$

This states the compatibility relation (1.45). Hence, we have

$$\nabla^2 \mathsf{u}(\underline{u}(\underline{v})) \cdot \nabla \underline{u}(\underline{v}) \quad = \quad \mathbf{I},$$

and the matrix

$$\nabla^2 \mathsf{u}(\underline{u}(\underline{v})) \quad = \quad (\nabla \underline{u}(\underline{v}))^{-1}$$

is positive definite. This proves that u is a strictly convex entropy. ∎

Entropy solutions

Due to their nonlinearity conservation laws tend to develop discontinuous solution from smooth initial data in finite time. Therefore, a new form of solution has to be considered. This is done by the concept of weak solutions, which describes the solution in a distributional sense.

Classical solutions of the Cauchy problem

We consider the Cauchy problem for systems of the form (1.26)

$$\partial_t \underline{u} + \sum_{j=1}^{p} \partial_{x_j} \underline{f}_j(\underline{u}) \quad = \quad 0, \tag{1.48}$$

$$\underline{u}(\underline{x}, 0) \quad = \quad \underline{u}_0(\underline{x}). \tag{1.49}$$

A classical solution of (1.48),(1.49) is a C^1 solution for $t > 0$, continuous for $t \geq 0$, which satisfies (1.48),(1.49) pointwise. When $\underline{u}_0$ is also of class C^1, it is a classical solution for $t \geq 0$.

Weak solutions

As in the scalar case we use the concept of weak solutions to have a more general solution form in mind which admits discontinuities. Classical solutions are not general enough to resolve such solutions for (1.26). Therefore, we consider $\underline{u}$ and $\underline{f}_i(\underline{u})$ in the distributional sense. Being more precise (and general) $\underline{u}$ must be supposed as a vector-valued locally bounded measurable function, i.e. $\underline{u} \in L^\infty_{\mathsf{loc}}([0, +\infty[\times\mathbb{R}^d)^p$, so that the $\underline{f}_i$ are defined pointwise.

Doing so and applying similar techniques as in the foregoing section, now for vector-valued functions, we derive the weak solution for the system of conservation laws (1.26) as

$$\int_0^\infty \int_{\mathbb{R}^d} \left(\langle \underline{u}, \partial_t \underline{\phi} \rangle + \sum_{i=1}^{d} \langle \underline{f}_i, \partial_{x_i} \underline{\phi} \rangle \right) d\underline{x}\, dt + \int_{\mathbb{R}^d} \langle \underline{u}_0(\underline{x}), \underline{\phi}(0, \underline{x}) \rangle \, d\underline{x} = 0.$$

Here, $\underline{\phi} \in C_0^1([0, +\infty[, \mathbb{R}^d)^p$ which means that the test functions $\underline{\phi}$ are the restriction to $([0, +\infty[\times\mathbb{R}^d) =: \Omega$ of C^1 functions with compact support in an open set containing Ω.

Entropy notion

We derive the entropy inequality for systems of conservation laws similar to the scalar case, i.e. by passing to the limit from a dissipative system. To this use, we follow the text by Godlewski & Raviart [31].

Considering a smooth solution $\underline{u}$ of (1.48), one might wonder whether $\underline{u}$ satisfies an additional conservation law of the form

$$\partial_t \mathsf{u}(\underline{u}) + \sum_{i=1}^{d} \partial_{x_i} \mathsf{f}_i(\underline{u}) \quad = \quad 0. \tag{1.50}$$

Here, u and f_i, $1 \leq i \leq d$ are sufficiently smooth functions, mapping from Ω to $\mathbb{R}$. This is indeed the case if they obey the following compatibility condition:

$$\mathsf{u}'(\underline{u})\underline{f}_i'(\underline{u}) \quad = \quad \mathsf{f}_i'(\underline{u}), \qquad 1 \leq i \leq d. \tag{1.51}$$

For ease of notations, $\mathsf{u}'(\underline{u}), \mathsf{f}_i'(\underline{u}) : \mathbb{R}^p \to \mathbb{R}^p$ are identified with the corresponding row vectors

$$\begin{aligned} \mathsf{u}' = \nabla_{\underline{u}} \mathsf{u}^T &= (\partial_{u_1} \mathsf{u}, \ldots, \partial_{u_p} \mathsf{u}), \\ \mathsf{f}_i' = \nabla_{\underline{u}} \mathsf{f}_i^T &= (\partial_{u_1} \mathsf{f}_i, \ldots, \partial_{u_p} \mathsf{f}_i). \end{aligned}$$

The linear mapping $\underline{f}_i' : \mathbb{R}^p \to \mathbb{R}^p$ is identified with the Jacobian

$$\underline{f}_i' = \mathbf{A}_i = \partial_{\underline{u}_k} f_i^j, \qquad \mathbf{A}_i \in \mathbb{R}^+ \times \mathbb{R}^p, \qquad 1 \leq j, k \leq p.$$

Since $\underline{u}$ is assumed as a classical solution, one can carry out the differentiation, multiply (1.48) by $\mathsf{u}'(\underline{u})$ and consider the compatibility condition (1.51) to obtain

$$\mathsf{u}'(\underline{u})\partial_t \underline{u} + \sum_{i=1}^{d} \mathsf{f}_i'(\underline{u})\partial_{x_i} \underline{u} \quad = \quad 0,$$

which is equal to (1.50).

Note that as in the scalar case, this is true for any classical solution of (1.48) but not in general for a weak solution and not in particular for a piecewise C^1 weak solution. In the following we examine several approaches to entropy formulations for conservation laws.

The viscous case

We have already seen in (1.21) that one way of introducing the framework of entropy solutions is to consider the limit of the dissipative case. For systems of conservation laws this is done in a similar manner by passing to the limit of a viscous system (cf. [66, 32]). Consider the hyperbolic system (1.48) in quasi linear form augmented with a perturbation, i.e. an artificial viscosity term:

$$\partial_t \underline{u}_\varepsilon + \sum_{i=1}^{d} f_i(\underline{u}_\varepsilon)\partial_{x_i} \underline{u}_\varepsilon \quad = \quad \varepsilon \Delta \underline{u}_\varepsilon, \qquad \varepsilon > 0. \tag{1.52}$$

Assume that the systems (1.52) augmented with initial data $\underline{u}_\varepsilon(0,\underline{x}) = \underline{u}_{0\varepsilon}(\underline{x}) \to \underline{u}_0(\underline{x})$ as $\varepsilon \searrow 0$ have sufficiently smooth solutions $\underline{u}_\varepsilon$.

Applying $\mathsf{u}'(\underline{u}_\varepsilon)$ to the dissipative system (1.52) gives

$$\mathsf{u}'(\underline{u}_\varepsilon)\partial_t \underline{u}_\varepsilon + \sum_{i=1}^{d} \mathsf{u}'(\underline{u}_\varepsilon)\underline{f}_i'(\underline{u}_\varepsilon)\partial_{x_i}\underline{u}_\varepsilon \quad = \quad \varepsilon \mathsf{u}'(\underline{u}_\varepsilon)\Delta \underline{u}_\varepsilon.$$

Using the compatibility relations (1.45) this is nothing more than

$$\partial_t \mathsf{u}(\underline{u}_\varepsilon) + \sum_{i=1}^{d} \partial_{x_i} \mathsf{f}_i'(\underline{u}_\varepsilon) \quad = \quad \varepsilon \mathsf{u}'(\underline{u}_\varepsilon)\Delta \underline{u}_\varepsilon.$$

Hence, the right-hand side of the system can be rewritten as

$$\varepsilon \mathsf{u}'(\underline{u}_\varepsilon)\Delta \underline{u}_\varepsilon \quad = \quad \varepsilon \Delta \mathsf{u}(\underline{u}_\varepsilon) - \varepsilon \sum_{i=1}^{d} (\partial_{x_i}\underline{u}_\varepsilon)^T \mathsf{u}''(\underline{u}_\varepsilon)\partial_{x_i}\underline{u}_\varepsilon,$$

and by the convexity of the entropy function u we have

$$\varepsilon \mathsf{u}'(\underline{u}_\varepsilon)\Delta \underline{u}_\varepsilon \quad \leq \quad \Delta \mathsf{u}(\underline{u}_\varepsilon).$$

Consequently, the entropy inequality reads as

$$\partial_t \mathsf{u}(\underline{u}_\varepsilon) + \sum_{i=1}^{d} \partial_{x_i} \mathsf{f}_i'(\underline{u}_\varepsilon) \quad \leq \quad \varepsilon \Delta \mathsf{u}(\underline{u}_\varepsilon). \tag{1.53}$$

If we assume $(\underline{u}_\varepsilon)_\varepsilon$ as a sequence of solutions of (1.52) bounded by a constant $C \geq 0$ independent of ε , i.e. $\|\underline{u}_\varepsilon\|_{L^\infty([0,+\infty[\times\mathbb{R}^d)^p} \leq C$, and converging almost everywhere to $\underline{u}$. Then

$$\underline{u}_\varepsilon \to \underline{u} \quad \text{as} \quad \varepsilon \to 0 \quad \text{in} \quad \mathcal{D}'(]0,+\infty[\times\mathbb{R}^d)^p,$$

i.e. in the sense of distributions on $(]0,+\infty[\times\mathbb{R}^d)^p$, so that

$$\partial_t \underline{u}_\varepsilon \to \partial_t \underline{u}, \quad \varepsilon \Delta \underline{u}_\varepsilon \to 0 \quad \text{in} \quad \mathcal{D}'(]0,+\infty[\times\mathbb{R}^d)^p.$$

With the above assumptions and the Lebesgue dominated convergence theorem, we have

$$\underline{f}_i(\underline{u}_\varepsilon) \to \underline{f}_i(\underline{u}) \quad \text{in} \quad L^1_{\mathsf{loc}}(]0,+\infty[\times\mathbb{R}^d)^p.$$

Passing (1.52) to the limit we have proven that $\underline{u}$ is a solution of (1.26) in the sense of distributions on $]0,+\infty[\times\mathbb{R}^d$.

With similar considerations we derive

$$\partial_t \mathsf{u}(\underline{u}_\varepsilon) \to \partial_t \mathsf{u}(\underline{u}), \quad \partial_{x_i}\mathsf{f}_i(\underline{u}_\varepsilon) \to \partial_{x_i}\mathsf{f}_i(\underline{u}), \quad \varepsilon\Delta \mathsf{u}(\underline{u}_\varepsilon) \to 0$$

in $\mathcal{D}'(]0,+\infty[\times\mathbb{R}^d)^p$.

Passing to the limit in (1.53) gives

$$\partial_t \mathsf{u}(\underline{u}_\varepsilon) + \sum_{i=1}^{d} \partial_{x_i} \mathsf{f}_i'(\underline{u}_\varepsilon) \quad \leq \quad 0.$$

Hence, we have proven the following

Theorem 1.36
Assume that (1.26) admits a convex entropy function u *with corresponding entropy fluxes* f_i, $1 \leq i \leq d$. *Let* $(\underline{u}_\varepsilon)_\varepsilon$ *be a sequence of sufficiently smooth solutions of (1.52) such that*

$$\|\underline{u}_\varepsilon\|_{L^\infty(]0,+\infty[\times\mathbb{R}^d)^p} \leq C,$$
$$\underline{u}_\varepsilon \to \underline{u} \quad \text{as} \quad \varepsilon \to 0 \quad \text{a.e. in} \quad]0,+\infty[\times\mathbb{R}^d,$$

where $C > 0$ *is a constant independent of* ε. *Then* $\underline{u}$ *is a weak solution of (1.26) and satisfies the entropy condition*

$$\partial_t \mathsf{u}(\underline{u}) + \sum_{i=1}^{d} \partial_{x_i} \mathsf{f}_i'(\underline{u}) \quad \leq \quad 0 \tag{1.54}$$

in the sense of distributions on $]0,+\infty[\times\mathbb{R}^d$.

Although this may seem a paradox, all exact science is dominated by the idea of approximation.

Bertrand Russell

2 Numerical approximations for conservation laws

This chapter is concerned with numerical approximation techniques for conservation laws, namely finite difference and finite volume schemes. Although there are several other possibilities, e.g. finite element or finite element related schemes like residual distribution schemes, we restrict ourselves for the sake of simplicity to this representation.

We describe basic techniques for scalar conservation laws in one space dimensions. Extensions to systems and several space dimensions are straight forward by dimensional splitting approaches.

A broader overview about numerical methods for conservation laws can be found in [31, 61, 71]. If no other references are mentioned, we refer to these books.

2.1 Basic concepts

We look for numerical approximations of the Cauchy problem for scalar conservation laws in one space dimensions, i.e.

$$\partial_t u + \partial_x f(u) \quad = \quad 0 \quad \text{in} \quad \mathbb{R}^+ \times \mathbb{R}, \quad u(0,\cdot) = u_0 \quad \text{in} \quad \mathbb{R}. \tag{2.1}$$

Assuming a uniform grid on $\mathbb{R}^+ \times \mathbb{R}$ with $x_i := i\Delta x$ and $t^n := n\Delta t$, the integral form of (2.1) for a spatial cell $C_i = [x_{i-1/2}, x_{i+1/2}]$ and a time interval $[t^n, t^{n+1}]$ can be written as

$$\int_{x_{i-1/2}}^{x_{i+1/2}} [u(t^{n+1},x) - u(t^n,x)]\,dx + \int_{t^n}^{t^{n+1}} [f(u(\xi,x_{i+1/2})) - f(u(\xi,x_{i-1/2}))]\,d\xi \quad = \quad 0. \tag{2.2}$$

According to the balance law (1.2) this reflects the physical principle of conservation. Thus, the change of the quantity u in the cell C_i during the time interval $[t^n, t^{n+1}]$ in the absence of sinks or sources is balanced by the flow difference through the cell faces $x_{i+1/2}$ resp. $x_{i-1/2}$.

If we consider approximations of the cell averages $\overline{u}(\cdot, t^n), \overline{u}(\cdot, t^{n+1})$ for the cell C_i, i.e.

$$\begin{aligned} U_i^n \approx \overline{u}(t^n, \cdot) &= \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} u(t^n, \xi)\, d\xi \\ U_i^{n+1} \approx \overline{u}(t^{n+1}, cdot) &= \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} u(t^{n+1}, \xi)\, d\xi, \end{aligned}$$

we are able to write (2.2) as

$$\begin{aligned} &\frac{1}{\Delta t}\left(U_i^{n+1} - U_i^n\right) \\ &+ \frac{1}{\Delta x_i}\left(\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f(u(\xi, x_{i+1/2}))\, d\xi - \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f(u(\xi, x_{i-1/2}))\, d\xi\right) = 0. \end{aligned} \tag{2.3}$$

Defining numerical approximations for the flux integrals,

$$\begin{aligned} F_{i+1/2}^n &:= F(U_{i-k+1}, \ldots, U_{i+k}) \approx \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f(u(\xi, x_{i+1/2}))\, d\xi \quad , \\ F_{i-1/2}^n &:= F(U_{i-k}, \ldots, U_{i+k-1}) \approx \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f(u(\xi, x_{i-1/2}))\, d\xi, \end{aligned} \tag{2.4}$$

the discrete analogue of (2.1),(2.2) is

$$\frac{1}{\Delta t}\left(U_i^{n+1} - U_i^n\right) = \frac{1}{\Delta x_i}\left(F_{i+1/2} - F_{i+1/2}\right). \tag{2.5}$$

We follow the notation of Sweby [104], i.e values with the index i as superscript denotes values at time t^{n+1}, those with subscript values at time t^n, i.e. $U_i^{n+1} = U^i$ resp. $U_i^n = U_i$. Thus, rearranging (2.5) to compute the cell average for time level t^{n+1} gives

$$U^i = U_i - \lambda[F_{i+1/2} - F_{i+1/2}], \tag{2.6}$$

with grid coefficient $\lambda := \Delta t/\Delta x_i$. This is an explicit numerical scheme to compute the cell averages $U(\cdot, t^{n+1})$ from known values at time level t^n. An implicit scheme will use approximations from both time levels to compute the unknowns at new time t^{n+1}, but such approximation classes will not be considered in this thesis.

A simple and accurate choice for the approximation of the flux function $f(u)$ is a central approximation, i.e. $F_{i+1/2} = \frac{1}{2}[f(U_{i+1}) + f(U_i)]$. Endowed with this choice for the numerical flux (2.4) the discretisation (2.3) is a second order approximation of (2.1) but unconditionally unstable (cf. [90]). Therefore, one major question in the following will be to find a proper approximation for the numerical flux functions (2.4).

An additional important question arising in the computation of discontinuous solutions of partial differential equations is the convergence of the numerical solution. How can one ensures convergence of the approximative solution to a weak solution of the conservation law? This question contains several difficulties, e.g. the numerical solution may converge to a genuine weak solution, however how can we ensure, that it converges to the physically relevant entropy-consistent solution.

One necessary condition to avoid convergence against an entropy violating weak solution is the conservative formulation of the numerical flux:

Definition 2.1
A finite volume scheme of the form (2.3) can be put in conservative form[1] *if there exists a continuous function* $F : \mathbb{R}^{2k} \to \mathbb{R}$ *such that*

$$\begin{aligned} \mathcal{H}(U_{i-k}, \ldots, U_{i+k}) &= U_i - \lambda \left[F(U_{i-k+1}, \ldots, U_{i+k}) - F(U_{i-k}, \ldots, U_{i+k-1}) \right], \\ &= U_i - \lambda \left[F_{i+1/2} - F_{i-1/2} \right]. \end{aligned} \tag{2.7}$$

F is called the numerical flux *and $F_{i+1/2} = F(U_{i+k}, \ldots, U_{i-k+1})$ denotes the flux between the cells C_i and C_{i+1} through the cell interface at $x_{i+1/2}$.*

Remark 2.2
It is obvious, that the conservative formulation of the numerical scheme reveals the conservation property of the partial differential equation. Assume that $U := (\ldots, U_{i-1}, U_i, U_{i+1}, \ldots) \in l^1(\mathbb{Z})$ and use the conservative formulation (2.3), the discrete solution operator $\mathcal{H}_\Delta$ maps a sequence U into the sequence $\mathcal{H}_\Delta(U)$, i.e.

$$\begin{array}{rccl} \mathcal{H}_\Delta : & l^1(\mathbb{Z}) & \to & l^1(\mathbb{Z}) \\ & U & \mapsto & \sum_{i\in\mathbb{Z}} \mathcal{H}_\Delta(U) = (\mathcal{H}_\Delta(U)_i)_{i\in\mathbb{Z}}) \end{array}, \tag{2.8}$$

with

$$\mathcal{H}_\Delta(U)_i = \mathcal{H}(U_{i-k}, \ldots, U_{i+k}).$$

The sum of this operator is

$$\begin{aligned} \sum_{i\in\mathbb{Z}} U^i &= \sum_{i\in\mathbb{Z}} \mathcal{H}(U_{i-k}, \ldots, U_{i+k}) \\ &= \sum_{i\in\mathbb{Z}} U_i - \lambda [F_{i+1/2} - F_{i-1/2}] \\ &= \sum_{i\in\mathbb{Z}} U_i - \lambda \sum_{i\in\mathbb{Z}} [F_{i+1/2} - F_{i-1/2}]. \end{aligned}$$

Since we have (for finite sums)

$$\sum_{i\in\mathbb{Z}} [F_{i+1/2} - F_{i-1/2}] = 0,$$

[1] This reveals the fact that the scheme is discretising the conservative form instead of the quasi linear one.

we obtain

$$\sum_{i\in\mathbb{Z}} U^i \quad = \quad \sum_{i\in\mathbb{Z}} U_i,$$

i.e. preservation of the discrete integral. Thereby, the discrete operator possesses the analogue property as the continuous solution operator (see [31] for details).

Naturally, the fundamental property of a conservative scheme (2.3) is the requirement of approximating the correct equation, i.e consistency with equation (2.1):

Definition 2.3
The scheme is said to be consistent *with (2.1), if*

$$F(u,\dots,u) \quad = \quad f(u), \quad \forall u \in \mathbb{R},$$

up to an additive constant $c \in \mathbb{R}$.

For a numerical scheme of the form (2.3) satisfying these fundamental requirements we have the well known Lax-Wendroff Theorem concerning the convergence of the numerical solution. In order to state this fundamental Theorem we need some notation. Let $(k_m)_m$ and $(h_m)_m$ be sequences converging to zero and $\Delta x = h_m, \Delta t = k_m = \lambda h_m$ with λ kept constant. For given initial values $u_0 \in L^1(\mathbb{R})$ we define cell averaged initial values

$$U_i^0 \quad := \quad \frac{1}{\Delta x_i} \int\limits_{x_{i-1/2}}^{x_{i+1/2}} u_0(x)\,dx. \tag{2.9}$$

After these preliminaries we are able to formulate the

Theorem 2.4 (Lax-Wendroff Theorem [69])
Let (2.7) be a conservative scheme consistent with equation (2.1) and initial values U^0 given by (2.9). Assume that there exists a sequence $(U)_m$ of discrete solutions with respect to h_m. Let h_m tend to zero and

i) $\|(U)_m\|_{L^\infty((0,+\infty)\times\mathbb{R})} \le C$,

ii) $(U)_m$ *converges in* $L^1_{loc}((0,+\infty)\times\mathbb{R})$ *and a.e. to a function* u.

Then u is a weak solution of (2.1).

Proof [69] ∎

Monotone schemes and TVD formulation

Monotone schemes are important for the numerical approximation of conservation laws since they build l^1-contracting semi-groups. This models, on a discrete level, the fact that no new extrema are created by a hyperbolic conservation law. Furthermore, for monotone schemes several theoretical results on there mathematical properties are known. Contrarily, for high-order approximations most of these results does not hold.

Monotone schemes

We start with the definition of a monotone approximation for (2.1):

Definition 2.5
A finite difference scheme of the form (2.7) is monotone *if the discrete solution operator (2.8) is a monotone increasing function of each argument, i.e.*

$$\partial_{U_i}\mathcal{H}(U_{i-k},\ldots,U_{i+k}) \geq 0, \qquad i-k \leq i \leq i+k.$$

Remark 2.6
An alternative formulation for monotonicity is the requirement that, if we write the scheme (2.7) in the form

$$\mathcal{H}(U_{i-k},\ldots,U_{i+k}) \;=\; \sum_{j=i-k}^{i+k} c_j U_j, \tag{2.10}$$

the coefficients c_j satisfy

$$c_j \geq 0, \quad \forall j \in \{i-k,\ldots,j+k\}. \tag{2.11}$$

Numerical methods which obey this property are called positive schemes.

Harten, Hyman and Lax [46] showed that a monotone scheme possesses the truncation error

$$\begin{aligned} u(t+\Delta t,x) - \mathcal{H}(u(t,x-k\Delta x),\ldots,u(t,x+k\Delta x)) \\ = -(\Delta t)^2\partial_x[\beta(u,\lambda)\partial_x u] + O((\Delta t)^3) \end{aligned}$$

where

$$\beta(u,\lambda) \;=\; \frac{1}{2\lambda^2}\sum_{j=-k}^{k} j^2 H_j(u,\ldots,u) - \frac{1}{2}(f'(u))^2,$$

i.e. models in fact the equation

$$\partial_t u + \partial_x f(u) \;=\; \partial_x(\beta(u)\partial_x u). \tag{2.12}$$

Except in the trivial case, we have $\beta \geq 0$ and $\beta \neq 0$ for monotone schemes. (2.12) is called the modified equation or first differential approximation of the numerical scheme (2.7). This states, that the scheme is a second-order accurate approximation to a viscous problem – the modified equation – which explains the diffusive nature of monotone methods. This shows that monotone schemes are at most first-order accurate. In the linear case, this result was already derived by Godunov and Lax [33, 65].

TVD formulation

Even though first-order methods possess many desirable properties like monotonicity, they are not very efficient in computing accurate numerical solutions. The monotonicity demand seems too restrictive to obtain high-order accurate methods.

In order to construct high-order schemes Harten[42] proposed a weaker stability condition given by total variation non-increasing methods, which ensure that the total variation

$$\mathsf{TV}(U(t^n,\cdot)) \;=\; \sum_{i\in\mathbb{Z}} |U_{i+1}-U_i|$$

is not increasing during advancing in time.

Remark 2.7
In the literature the slightly imprecise formulation total variation diminishing *rather than* total variation non-increasing *is used. In the remainder we follow this convention.*

Definition 2.8
A finite volume scheme of the form (2.7) is said to be

– total variation diminishing *(TVD) if*

$$\mathsf{TV}(\mathcal{H}(U)) \;\leq\; \mathsf{TV}(U), \tag{2.13}$$

– monotonicity preserving *if*

$$U^n \;\; \textit{monotone} \implies U^{n+1} \textit{ monotone}. \tag{2.14}$$

For the classes of numerical schemes we have already discussed, Harten [42] stated the following properties:

Proposition 2.9
i) A monotone scheme is TVD.

ii) A TVD scheme is monotonicity preserving

Proof

i) Harten, Hyman and Lax [46] proved that monotone schemes form a l_1-contractive semi-group, i.e.

$$\|\mathcal{H}(V)-\mathcal{H}(W)\|_{l_1} \leq \|V-W\|_{l_1}$$

for all l_1-summable function V and W with

$$\|U\|_{l_1} \;=\; \sum_{i=-\infty}^{\infty} |U_i|.$$

Choosing $V=U$ and $W=\mathcal{T}U$ where $\mathcal{T}$ is the translation operator (i.e. $W_i=U_{i+1}$), we get

$$\begin{aligned}\|\mathcal{H}(U)-\mathcal{H}(W)\|_{l_1} &= \|\mathcal{H}(U)-\mathcal{H}(\mathcal{T}U)\|_{l_1}\\ &= \mathsf{TV}(\mathcal{H}(U))\\ &\leq \mathsf{TV}(U)\end{aligned}$$

which proves i).

ii) We consider a sequence of the form

$$U = \begin{cases} U^- = \text{constant}, & \forall i \le I^- \\ \text{monotone} & I^- \le i \le I^+ \\ U^+ = \text{constant}, & \forall i \ge I^+. \end{cases}$$

Obviously the total variation of the sequence is $\mathsf{TV}(U) = |U^+ - U^-|$. If the solution on the new time level $V = \mathcal{H}(U)$ is not monotone, it possesses at least one local minimum U^m and one local maximum U^M. We get

$$\mathsf{TV}(V) \ge |U^+ - U^-| + |U^M - U^m| > \mathsf{TV}(U),$$

which is a contradiction to the assumption that the scheme is TVD.

■

Godunov proved that any linear monotonicity preserving scheme, and therefore any TVD scheme, is a monotone scheme and consequently of first-order accuracy. This does not exclude the possibility of having nonlinear TVD schemes which are second order accurate. To this purpose, Harten [42] introduced the increment notion for three-point and five-point schemes.

Incremental form and numerical viscosity

Definition 2.10
The scheme (2.7) is said to be in increment form*, if there exist two functions of $2k$ variables C, D called incremental coefficients,*

$$C_{i+1/2} := C(U_{i-k+1}, \ldots, U_{i+k}), \tag{2.15}$$
$$D_{i+1/2} := D(U_{i-k+1}, \ldots, U_{i+k}), \tag{2.16}$$

such that we can write the scheme as

$$\mathcal{H}(U_{i-k}, \ldots, U_{i+k}) = U_i + C_{i+1/2}\Delta U_{i+1/2} - D_{i-1/2}\Delta U_{i-1/2}, \tag{2.17}$$

with $\Delta U_{i+1/2} := U_{i+1} - U_i$..

Proposition 2.11
Any consistent conservative three-point scheme of the form (2.7) with Lipschitz continuous numerical flux F admits a unique incremental form with incremental coefficients given by

$$\begin{aligned} C_{i+1/2} &= \lambda \frac{[f_i - F_{i+1/2}]}{\Delta U_{i+1/2}}, \\ D_{i+1/2} &= \lambda \frac{[f_{i+1} - F_{i+1/2}]}{\Delta U_{i+1/2}}. \end{aligned} \tag{2.18}$$

Lemma 2.12
If the coefficients (2.15) satisfy the inequalities

$$\begin{aligned} C_{i+1/2} &\geq 0, \\ D_{i-1/2} &\geq 0, \\ C_{i+1/2} + D_{i-1/2} &\leq 1, \end{aligned} \tag{2.19}$$

the scheme (2.17) is TVD.

Proof [42] ∎

It will be useful to consider the following class of schemes:

Definition 2.13
A numerical scheme of the form (2.7) is called essentially three-point *if its numerical flux satisfies the stronger consistency relation*

$$F(U_{i-k+1}, \ldots, U_{i-1}, u, u, U_{i+2}, \ldots, U_{i+k}) = f(u). \tag{2.20}$$

For essentially three-point schemes we can give the following characterisation:

Definition 2.14
A numerical scheme of the form (2.3) is said to be in viscosity form*, if there exists a function Q of $2k$ variables called the* numerical viscosity coefficient*,*

$$Q_{i+1/2} := Q(U_{i-k+1}, \ldots, U_{i+k}),$$

such that we can write it as

$$\begin{aligned} \mathcal{H}(U_{i-k}, \ldots, U_{i+k}) = U_i - \frac{\lambda}{2} \big[&(f_{i+1} - f_{i-1}) \\ &+ Q_{i+1/2} \Delta U_{i+1/2} - Q_{i-1/2} \Delta U_{i-1/2} \big]. \end{aligned} \tag{2.21}$$

Thus, we can write the numerical scheme as a combination of a central difference of the flux function – which is unconditionally unstable (see [90]) – and a viscous term to stabilise the method. Consequently the numerical flux reads as

$$F_{i+1/2} = \frac{1}{2}(f_{i+1} + f_i) - \frac{1}{2\lambda} Q_{i+1/2} \Delta U_{i+1/2}.$$

Hence, the scheme (2.21) is an essentially three-point one. This observation leads to the following Lemma proved by Tadmor [107] which builds a bridge to the unique representation of schemes by their incremental coefficients.

Lemma 2.15
Any essentially three-point scheme admits a unique representation by their viscous form. The coefficient of numerical viscosity is given by

$$Q_{i+1/2} = \frac{[f_{i+1} + f_i - 2F_{i+1/2}]}{\Delta U_{i+1/2}}.$$

This coefficient can be expressed by the unique incremental coefficients $C_{i+1/2}, D_{i+1/2}$ of a three-point scheme as

$$Q_{i+1/2} = C_{i+1/2} + D_{i+1/2}. \tag{2.22}$$

Here, we see the need to distinguish between essentially three-point schemes and three-point schemes. In general the incremental coefficients of an essentially three-point scheme are not defined in a unique way. Nevertheless they can be defined in a similar way. Using (2.18), (2.21), they can be reformulated as

$$\begin{aligned} C_{i+1/2} &= -\frac{1}{2}\left[\lambda\frac{\Delta f_{i+1/2}}{\Delta U_{i+1/2}} - Q_{i+1/2}\right], \\ D_{i+1/2} &= \frac{1}{2}\left[\lambda\frac{\Delta f_{i+1/2}}{\Delta U_{i+1/2}} + Q_{i+1/2}\right]. \end{aligned}$$

We have seen that under conditions (2.15) a scheme written in incremental form is TVD. Because of the ability to express the incremental coefficients by the numerical viscosity coefficient of a scheme and vice versa, it is clear that we have equivalent inequalities for (2.15) to ensure that a scheme in viscous form is TVD:

Corollary 2.16
Assume a numerical scheme (2.7) written in viscous form (2.21). Let the numerical viscosity coefficient satisfies

$$\lambda\left|\frac{\Delta f_{i+1/2}}{\Delta U_{i+1/2}}\right| \le Q_{i+1/2} \le 1 \qquad \forall i \in \mathbb{Z}.$$

Then the method is TVD.

Examples of some classical schemes

In the following we give some short examples of the most common and meanwhile 'classical' schemes for the numerical treatment of conservation laws. All these schemes are three point schemes. We will see in the following that schemes of this form possess some remarkable properties.

Example 2.17 (The upwind scheme [18, 84])
If we look back to the linear advection equation, the solution propagates in time along the characteristic curves where the solution is constant. So, if we trace back the solution along the characteristics given by $x - at$ over one cell, the solution at the point (t^{n+1}, x_i) is determined by the solution at the point $(t^n, x_i - a\Delta t)$, i.e.

$$\begin{aligned} u(t^{n+1}, x_i) &= u(t^n, x_i - a\Delta t) \\ &= u(t^n, x_i - a\lambda h). \end{aligned}$$

If we assume $a > 0$, i.e. the advection propagates to the right, it is clear that the point $(t^n, x_i - ah)$ is located on the left side of point (t^n, x_i). Since the information about the solution approaches from there, it is natural to compute the approximation of u at point $(t^n, x_i - ah)$

from a linear interpolation between the values of $u(t^n, x_{i-1})$ *and* $u(t^n, x_i)$*. Consequently we get*

$$\begin{aligned} u(t^{n+1}, x_i) &\approx \lambda a u(t^n, x_{i-1}) + (1-\lambda) u(t^n, x_i) \\ &= u(t^n, x_i) - \lambda a [u(t^n, x_i) - u(t^n, x_{i-1})]. \end{aligned}$$

Hence, the upwind idea is to use the information from the side where it comes from. This leads to the fact that for $a < 0$ *we have to take the node on the right to approximate* $u(t^{n+1}, x_i,)$*. The corresponding numerical scheme, the upwind scheme, reads as*

$$U^i = \begin{cases} U_i - \lambda a [U_i - U_{i-1}], & a \geq 0, \\ U_i - \lambda a [U_i - U_{i+1}], & a < 0, \end{cases}$$

or in a more compact form

$$U^i = U_i - a\frac{\lambda}{2}[U_{i+1} - U_{i-1}] + |a|\frac{\lambda}{2}[U_{i+1} - 2U_i + U_{i-1}].$$

The natural extension to the nonlinear case in this form is given by

$$U^i = U_i - \frac{\lambda}{2}[f_{i+1} - f_{i-1}] + \frac{\lambda}{2}[|a_{i+1/2}|(U_{i+1} - U_i) - |a_{i-1/2}|(U_i - U_{i-1})], \quad (2.23)$$

i.e.

$$Q^{up}_{i+1/2} = \lambda |a_{i+1/2}|, \quad (2.24)$$

with

$$a_{i+1/2} := \begin{cases} \left| \frac{f_{i+1} - f_i}{U_{i+1} - U_i} \right|, & U_{i+1} \neq U_i, \\ & \text{if} \\ f'(U_i), & U_{i+1} = U_i. \end{cases}$$

The first approach to schemes considering the directions of the characteristics and so the origin of the construction of upwind schemes was given by Courant, Isaacson and Rees in [18]. Their formulation is based on a quasi-linear system and is non-conservative. The conservative formulation (2.23),(2.24) was given by Murman [84]. This scheme is also known as Murman-Courant-Isaacson-Rees(MCIR) scheme.

The CFL condition

As one can easily see the slope of the characteristics determines the range of dependence. If we think of the advection equation (1.6), the solution (1.8) propagates along the ray of the characteristics, i.e. depends on the initial value traced back along these lines. For a wave equation which propagates in all spatial directions, the solution depends on the data inside of this cone. This area is called the domain of dependence of the partial differential equation (cf. [57]).

This fact imposes a geometrical condition on the ratio of the mesh size. One has to choose the time-step in such a way that the numerical domain of dependence (i.e. the stencil of

the numerical scheme) has to be enclosed in the mathematical domain of dependence of the solution. This reveals the fact that a numerical method only remains stable, if we use the data on which the solution really depends on, to approximate the solution (see e.g. [90] for details).

This restriction of the ratio of time and spatial mesh-size is given by the condition

$$\frac{\Delta t}{\Delta x}\max_{u\in U}|f'(u)| \;=\; \lambda\max_{u\in U}|f'(u)| \;\;\leq\;\; 1. \tag{2.25}$$

Here, one has to remember that f' reflects the inverse slope of the characteristics. (2.25) is known as the Courant-Friedrichs-Lewy[2](CFL) condition. This condition was derived in the fundamental work[16], where the first approach concerning stability and convergence for the numerical solution of partial differential equation was given. The number of the left-hand side on (2.25) is called Courant number or CFL number.

Example 2.18 (The Lax-Friedrichs scheme [64])
The Lax-Friedrichs scheme was the starting point for designing numerical methods for conservation laws. The original formulation of the scheme reads as

$$U^i = \frac{1}{2}(U_{i+1}+U_{i-1}) - \frac{1}{2}\lambda\left[f(U_{i+1}) - f(U_{i-1})\right]. \tag{2.26}$$

If we cast this formula into a different form, i.e.

$$\begin{aligned}\mathcal{H}^{LF} &= U_i - \frac{1}{2}\lambda\left[f(U_{i+1}) - f(U_{i-1})\right] + \frac{1}{2}(U_{i+1} - 2U_i + U_{i-1})\\ &= U_i - \lambda\left[\frac{1}{2}(f_{i+1}+f_i) - \frac{1}{2\lambda}(U_{i+1}-U_i)\right.\\ &\qquad\left. - \frac{1}{2}(f_i+f_{i-1}) + \frac{1}{2\lambda}(U_i - U_{i-1})\right],\end{aligned} \tag{2.27}$$

it is easily seen that the method is stabilised by a linear diffusion term, with constant dissipation coefficient

$$Q^{LF}_{i+1/2} \;=\; 1.$$

It is obvious that the last term in the first line of (2.27) is the discretisation of a viscous term. This explains the robust behaviour and the diffusivity of the Lax-Friedrichs scheme.

If we examine whether the LF-scheme is monotone, we compute

$$\begin{aligned}\partial_{U_{i+1}}\mathcal{H}^{LF} &= \frac{1}{2} - \frac{1}{2}\lambda f'(U_{i+1})\\ &= \frac{1}{2}[1-\lambda f'(U_{i+1})],\\ \partial_{U_i}\mathcal{H}^{LF} &= 0,\\ \partial_{U_{i-1}}\mathcal{H}^{LF} &= \frac{1}{2} - \frac{1}{2}\lambda f'(U_{i-1})\\ &= \frac{1}{2}[1-\lambda f'(U_{i-1})].\end{aligned}$$

[2]Richard Courant, Göttingen, New York (1888 – 1972),
Kurt-Otto Friedrichs, Göttingen, Braunschweig, New York (1901 – 1982),
Hans Lewy, Göttingen, Berkeley (1903 – 1988).

Since the CFL-condition guarantees that $\lambda f'(u) \leq 1$, $\forall u \in \mathbb{R}$, the scheme is monotone.

Example 2.19 (The Godunov scheme [33])
The Godunov scheme is based on the fact, that we can view the numerical solution between two cells as a local Riemann problem. Thus, we have to solve the Riemann problem at the cell boundaries $x_{i-1/2}$ and $x_{i+1/2}$ exactly from the piecewise constant data on C_{i-1}, C_i and C_{i+1}. In the end we have to project the propagated solution on the interval at time t^{n+1}.

The initial data for the local Riemann problem – centred at (t_0, x_0) – is

$$v_R(t_0, x; u_r, u_l) \;=\; \begin{cases} u_l, & x \leq x_0 \\ u_r, & x > x_0 \end{cases} \qquad \text{for} \quad x_l \leq x \leq x_r,$$

and is solved on the time interval $[t^n, t^{n+1}]$ exactly. The solution is self-similar, i.e. depends only on $\xi = (x - x_0)/(t - t_0)$ and the initial data (2.28), i.e.

$$u_R(t, x; u_r, u_l) \;=\; u_R(\xi; u_r, u_l).$$

The solution requires the stronger CFL condition

$$\lambda \max_{u \in U} |a(u)| \leq 1/2.$$

This reveals the fact that we have to choose the time step small enough, such that waves originating from different Riemann problems do not interact. The solution of the Riemann problem remains constant in the cell faces $x_{i+1/2}, x_{i-1/2}$ during the evolution in time, because of

$$u_R(t, x) \;=\; u\left(\frac{x - x_{i\pm 1/2}}{t}\right) \quad \text{for} \quad t \in (t^n, t^{n+1}], x \in (x_{i-1/2}, x_{i+1/2}].$$

Thus, we write

$$\begin{aligned} u_R(t^{n+1}, x_{i+1/2}; U_{i+1}, U_i)) \;&=\; u_R(0; U_{i+1}, U_i) \;=\; u_R^+ \\ u_R(t^{n+1}, x_{i-1/2}; U_i, U_{i-1})) \;&=\; u_R(0; U_i, U_{i-1}) \;=\; u_R^- \end{aligned}$$

and considering the integral form we get

$$\int_{x_{i-1/2}}^{x_{i+1/2}} u(t^{n+1}, x)\, dx \;=\; \int_{x_{i-1/2}}^{x_{i+1/2}} u(t^n, x)\, dx - \int_{t^n}^{t^{n+1}} \left[f(u(\xi, x_{i+1/2})) - f(u(\xi, x_{i-1/2}))\right] d\xi$$

which is equivalent to

$$\Delta x U^{i+1} \;=\; \Delta x U_i - \Delta t \left[f(u_R^+) - f(u_R^-)\right].$$

In the scalar case, for a convex flux function the Godunov scheme takes a very simple form. The Riemann problem writes as

$$\partial_t u + \partial_x f(u) \;=\; 0,$$

$$u^R(0;U_i,U_{i+1}) \;=\; \begin{cases} U_i & f'_{i+1/2} > 0 \\ & \text{for} \qquad\qquad x \in [x_i, x_{i+1}]\,, \\ U_{i+1} & f'_{i+1/2} < 0 \end{cases}$$

and so

$$F^G(U_i,U_{i+1}) \;=\; f(u^R(0;U_i,U_{i+1})) \;=\; \begin{cases} f(U_i) & f'_{i+1/2} > 0 \\ & \text{for} \qquad\qquad x \in [x_i, x_{i+1}] \\ f(U_{i+1}) & f'_{i+1/2} < 0 \end{cases} \tag{2.28}$$

which is the upwind scheme.

Remark 2.20
For a general flux function f there is still a simple expression for the numerical flux function F^G:

$$F^G(U_{i+1},U_i) \;=\; \begin{cases} \min\limits_{u^\star\in[U_i,U_{i+1}]} f(u^\star) & U_i \le U_{i+1} \\ & \text{for} \\ \max\limits_{u^\star\in[U_{i+1},U_i]} f(u^\star) & U_i > U_{i+1}, \end{cases} \tag{2.29}$$

(cf. [85, 59]).

From this definition it is straightforward to derive the viscosity coefficient for the Godunov scheme:

$$Q^G_{i+1/2} \;=\; \lambda \max_{(u-U_i)(u-U_{i+1})\le 0} \frac{f_{i+1}+f_i-2f(u)}{U_{i+1}-U_i}. \tag{2.30}$$

Approximative Riemann solver

It is clear that the computational cost for solving a Riemann problem for each cell might be quite large for complex applications. Thus, in the beginning of the eighties, the question arose if one could approximate the solution of a Riemann problem to overcome this difficulty.

Example 2.21 (The Roe scheme [92])
Roe proposed an approximative Riemann solver based on the upwind scheme, which gives a natural extension to nonlinear equations of this standard scheme. He approximated the original Riemann problem (2.28) by the linearised version

$$\partial_t u + \hat{a}(U_{i+1},U_i)\partial_x u \;=\; 0,$$

$$u^R(x,0) \;=\; \begin{cases} U_i, & x \le x_{i+1/2} \\ U_{i+1} & x > x_{i+1/2} \end{cases} \qquad \text{for } x_i \le x \le x_{i+1}.$$

Here, $\hat{a}(U_{i+1},U_i)$ is an approximation to the derivative of the flux function. The approximative Riemann solution $\hat{u}$ is a discontinuity wave propagating with speed $\hat{a}(U_{i+1},U_i)$, i.e.

$$\hat{u}(U_{i+1},U_i) = u_R(\xi,U_{i+1},U_i) \;=\; \begin{cases} U_i & \xi < \hat{a}(U_{i+1},U_i) \\ & \text{for} \\ U_{i+1} & \xi > \hat{a}(U_{i+1},U_i) \end{cases}$$

This leads to the approximative Godunov method

$$U^i \;=\; U_i - \lambda[f(\hat{u}(U_{i+1},U_i)) - f(\hat{u}(U_i,U_{i-1}))].$$

For the scalar case $\hat{a}$ is uniquely defined by

$$\hat{a}(U_{i+1},U_i) \;=\; \frac{f(U_{i+1})-f(U_i)}{U_{i+1}-U_i}.$$

Thereby, the flux reduces to (2.28) and is equivalent to the upwind scheme (2.23).

Remark 2.22
The extension to systems of equations is quite naturally and can also be found in [92] where Roe gives the following conditions for the approximative matrix $\hat{\mathbf{A}}(\underline{U}_{i+1},\underline{U}_i)$:

i) $\hat{\mathbf{A}}(\underline{U}_{i+1},\underline{U}_i)(U_{i+1}-U_i) = \underline{f}(U_{i+1}) - \underline{f}(U_i)$,

ii) $\hat{\mathbf{A}}(\underline{U}_{i+1},\underline{U}_i)$ *is diagonalisable with real eigenvalues,*

iii) $\hat{\mathbf{A}}(\underline{U}_{i+1},\underline{U}_i) \to \underline{f}'(\underline{u})$ *smoothly as* $\underline{U}_{i+1},\underline{U}_i \to \underline{u}$.

Unfortunately, the Roe scheme admits non-physical weak solutions. This drawback can be removed by an entropy fix proposed by Harten and Hyman [45].

Many similar approaches originating from approximative Riemann solver were made at the same time or short after Roe's suggestion. We only name some of them for completeness, e.g. the HLL-solver [47] and the extension (HLLE) by Einfeldt [24] and the Enquist-Osher scheme [26, 27]

Example 2.23 (modified Lax-Friedrichs scheme [105])
The method

$$\mathcal{H}^{LFm}(U_{i+1},U_i,U_{i-1}) \;=\; (U^{m+}_{i+1/2} + U^{m-}_{i-1/2})/2 \tag{2.31}$$

with

$$U^{m+}_{i+1/2} \;:=\; \frac{2}{\Delta x}\int_{x_i}^{x_{i+1}} u_R(\xi/\Delta t;U_{i+1},U_i)\,d\xi \;=\; (U_{i+1}+U_i)/2 - \lambda(f_{i+1}-f_i)$$

$$U^{m-}_{i-1/2} \;:=\; \frac{2}{\Delta x}\int_{x_{i-1}}^{x_i} u_R(\xi/\Delta t;U_i,U_{i-1})\,d\xi \;=\; (U_i+U_{i-1})/2 - \lambda(f_i-f_{i-1})$$

is called the modified Lax-Friedrichs scheme.

The numerical flux function can be cast in the form (2.21) with numerical dissipation coefficient

$$Q^{LFm}_{i+1/2} \;=\; \frac{1}{2}. \tag{2.32}$$

Lemma 2.24
Under the CFL condition

$$\lambda \max_{u\in\hat{u}} |f'(u)| \leq \frac{1}{2}$$

the modified Lax-Friedrichs scheme is monotone.

The motivation for the modification of the Lax-Friedrichs scheme originates in the comparison with the Godunov scheme, i.e. the possibility to write both in a similar form. Later we will see that this leads to a classification for entropy-stable schemes which can be written (with this notation) as a convex combination of the Godunov and the modified Lax-Friedrichs schemes. This approach and the modification of the Lax-Friedrichs scheme were given by Tadmor [105, 106].

Example 2.25 (The Lax-Wendroff scheme [70])
The Lax-Wendroff scheme is the only one among these examples which is second-order accurate. It was proposed by Lax and Wendroff [70] using Taylor series expansion not only in space but also in time. This is done in the following manner:

$$\begin{aligned} u(x,t+\Delta t) &= u(x,t) + \Delta t u_t(x,t) + \tfrac{1}{2}(\Delta t)^2 u_{tt} + O((\Delta t)^3) \\ &= u(x,t) + \Delta t[-f(u)_x] + \tfrac{1}{2}(\Delta t)^2 \partial_t[-f(u)_x] + O((\Delta t)^3) \\ &= u(x,t) - \Delta t f(u)_x - \tfrac{1}{2}(\Delta t)^2 \partial_x[f(u)_t] + O((\Delta t)^3) \\ &= u(x,t) - \Delta t f(u)_x - \tfrac{1}{2}(\Delta t)^2 \partial_x[Df(u)^2 u_x] + O((\Delta t)^3). \end{aligned}$$

If we now plug-in the numerical approximation U_i of the value $u(t,x_i)$ and use second-order approximations for the flux derivatives, we end up with the following formula:

$$\begin{aligned} U^i &= U_i - \lambda \Big[f_{i+1/2} - \frac{1}{2}\lambda a^2_{i+1/2}(U_{i+1} - U_i) \\ &\qquad - f_{i-1/2} + \frac{1}{2}\lambda a^2_{i+1/2}(U_i - U_{i-1}) \Big] + O((\Delta t)^3). \end{aligned} \tag{2.33}$$

Hence, one sees immediately that the viscosity coefficient for the Lax-Wendroff scheme reads as

$$Q^{LW}_{i+1/2} = \lambda^2 a^2_{i+1/2}.$$

Unfortunately the Lax-Wendroff fails to satisfy the fundamental concepts of monotony and TVD. In the vicinity of discontinuities the method produces spurious oscillations.

2.2 Entropy solutions

We already have seen that conservation laws satisfy in addition an entropy condition. Naturally, one wishes to model this property of the underlying equations inside the numerical scheme. The Lax-Wendroff Theorem guarantees that a convergent numerical scheme converges to a weak solution of the conservation law. Since an entropy inequality guarantees

the convergence to physically relevant solutions, one wishes to construct a numerical scheme obeying some nonlinear stability criteria using a discrete version of an entropy inequality (1.24).

In the following we will examine several conditions for scalar schemes in order to satisfy such entropy conditions.

Consistency

We start from the entropy inequality (1.24) for scalar conservation laws (2.1) with a convex entropy function u and the corresponding entropy flux $\mathsf{f}(u)$ related to the flux function $f(u)$ by the compatibility relation (1.22).

Definition 2.26
A difference scheme of the form (2.3) is consistent with the entropy condition

$$\partial_t \mathsf{u}(u) + \partial_x \mathsf{f}(u) \quad \leq \quad 0,$$

if there exists a continuous function $\mathsf{F} : \mathbb{R}^{2k} \to \mathbb{R}$ *which*

i) is consistent with the entropy flux f

$$\mathsf{F}(u, \ldots, u) = \mathsf{f}(u),$$

ii) satisfies a discrete entropy inequality

$$\begin{aligned} \mathsf{u}(U^i) - \mathsf{u}(U_i) &= \mathsf{u}^i - \mathsf{u}_i \\ &\leq \lambda \left[\mathsf{F}(U_{i-k+1}, \ldots, U_{i+k}) - \mathsf{F}(U_{i-k}, \ldots, U_{i+k-1}) \right]. \end{aligned} \tag{2.34}$$

u_i *is the the numerical entropy function for* U_i *and* F *the* numerical entropy flux.

As mentioned above the fundamental relevance of schemes satisfying a discrete entropy inequality is given through the following Theorem due to Harten, Hyman and Lax:

Theorem 2.27
Suppose that the conditions of the Lax-Wendroff Theorem – (Theorem 2.4) – hold and let the scheme be consistent with any entropy condition. Then the limit u *representing the weak solution of (1.14) is the unique entropy solution of (2.1).*

Proof It is obvious, using the techniques of the proof of the Lax-Wendroff Theorem, that if the difference scheme satisfies a discrete entropy inequality (2.34), the limit u satisfies the corresponding continuous entropy inequality (1.24) (see [46]). ∎

In the last section we have seen, that the demand of monotonicity is a strong condition for a numerical scheme. It implies the TVD and l_1-contractive properties as well as, at least in the linear case, first-order accuracy. The following theorem reveals the fact that monotonicity is an even stronger condition:

Theorem 2.28
A monotone consistent scheme is consistent with any entropy condition.

Proof The proof was given by Harten, Hyman and Lax in [46] while a different approach was presented by Crandall and Majda [19]. ∎

Remark 2.29
As a consequence of this considerations Schonbeck [94] proved that the Lax-Wendroff scheme (2.33) is not entropy stable. This means that the scheme does not fulfil the discrete entropy inequality (2.34), regardless what numerical entropy flux is used.

E-schemes

Osher and Tadmor developed a whole theory about schemes automatically obeying an entropy condition, called E-schemes. We do not want to repeat the whole development of this class of schemes. Here, we just cite some important results in order to present the assets as well as the drawbacks of these algorithms.

We start with the definition of E-schemes given by Osher [85].

Definition 2.30
A consistent conservative scheme is called an E-scheme *if its numerical flux satisfies*

$$sign(U_{i+1} - U_i)[F_{i+1/2} - f(u)] \leq 0, \quad \forall u \in [U_i, U_{i+1}].$$

From Theorem 2.28 concerning monotone schemes the following proposition is quite obvious:

Proposition 2.31
A three-point monotone scheme is an E-scheme.

Proof Since $F(u, v)$ is non-decreasing in u and non-increasing in v, one sees easily

$$\begin{aligned} F(u,v) \;&\leq\; F(u,w) \;\leq\; F(w,w) \quad \text{if } u \leq v \leq w, \\ F(u,v) \;&\geq\; F(w,v) \;\geq\; F(w,w) \quad \text{if } u \geq v \geq w, \end{aligned}$$

and obtains immediately

$$\text{sign}(v-u)[F(u,v) - F(w,w)] \leq 0, \quad \forall w \in [u, v].$$

∎

E-schemes can also be characterised by their dissipation coefficients. If one chooses the right bounds Tadmor [106] showed that they are consistent with any entropy condition:

Lemma 2.32
Assume that the CFL-like condition

$$\lambda \max_i \left| \frac{\Delta_{i+1/2} f}{\Delta_{i+1/2} u} \right| \leq 1$$

holds. Then the E-fluxes are characterised by

$$\begin{array}{rcll} F_{i+1/2} & \leq & F^G_{i+1/2} & U_i < U_{i+1}, \\ & & \textit{for} & \\ F_{i+1/2} & \geq & F^G_{i+1/2} & U_i \geq U_{i+1}, \end{array}$$

where F^G is the Godunov-flux.

Proof From (2.29) we have under CFL ≤ 1:

$$\begin{array}{ll} \min\limits_{u\in[U_i,U_{i+1}]} f(u) & U_i < U_{i+1}, \\ & \text{for} \\ \max\limits_{u\in[U_{i+1},U_i]} f(u) & U_i \geq U_{i+1}, \end{array}$$

This proves that the numerical flux of Godunov's method is the limit of the flux of an E-scheme. ∎

Theorem 2.33
Under the CFL condition

$$\lambda \max|a(u)| \leq 1/2,$$

an E-scheme with numerical dissipation coefficient $Q^E_{i+1/2}$, satisfying

$$Q^G_{i+1/2} \leq Q^E_{i+1/2} \leq Q^{mLF}_{i+1/2},$$

is consistent with any entropy condition.

Proof [106] ∎

Proposition 2.34
An E-scheme with differentiable numerical flux is at most first order accurate.

Proof [106] ∎

Discrete entropy inequalities

As already mentioned, there are some degrees of freedom to choose the numerical entropy flux because the only requirement is the compatibility condition (1.22). So we are faced with the problem that every numerical entropy flux of the form

$$\mathsf{F}(u,v) \;:=\; \frac{1}{2}[\mathsf{f}(u)+\mathsf{f}(v)] - \frac{1}{2\lambda}\mathsf{Q}(u,v), \quad \mathsf{Q}(u,u)=0 \quad \forall u$$

is a consistent choice.

Numerical entropy flux

Lax [66] was the first who proved entropy stability for the Lax-Friedrichs scheme (2.27). He applied the scheme to a system of conservation laws and showed that it satisfies a discrete entropy inequality with numerical entropy flux chosen as

$$\mathsf{F}(U_{i+1},U_i) \;=\; \frac{1}{2}[\mathsf{f}(U_{i+1})+\mathsf{f}(U_i)]-\frac{1}{2\lambda}[\mathsf{u}(U_{i+1})-\mathsf{u}(U_i)].$$

Exploiting this assumption Sonar [102] proposed the following

Definition 2.35
A numerical entropy flux is called Lax-consistent *if it can be written in the form*

$$\begin{aligned}\mathsf{F}_{i+1/2} &= \mathsf{F}(U_{i+1},U_i)\\ &= \frac{1}{2}[\mathsf{f}(U_{i+1})+\mathsf{f}(U_i)]-\frac{1}{2\lambda}\mathsf{Q}(U_{i+1},U_i)[\mathsf{u}(U_{i+1})-\mathsf{u}(U_i)]\\ &= \frac{1}{2}[\mathsf{f}_{i+1}+\mathsf{f}_i]-\frac{1}{2\lambda}\mathsf{Q}(U_{i+1},U_i)[\mathsf{u}_{i+1}-\mathsf{u}_i]\end{aligned}$$

with numerical entropy dissipation coefficient Q.

Sonar proposed a numerical entropy flux of the form

$$\begin{aligned}\mathsf{F}_{i+1/2} &= \mathsf{F}(U_{i+1},U_i)\\ &= \frac{1}{2}[\mathsf{f}(U_{i+1})+\mathsf{f}(U_i)]-\frac{1}{2\lambda}\mathsf{Q}(U_{i+1},U_i)[\mathsf{u}(U_{i+1})-\mathsf{u}(U_i)].\end{aligned}$$

where Q stems from Q by replacing u, f and the corresponding derivations by u,f and their derivations. He has to show c-consistency (compatibility-consistency) for the scheme:

Definition 2.36
A consistent numerical entropy flux F *corresponding to a numerical flux* F*, satisfying*

$$\begin{aligned}\mathsf{u}'(v)\partial_u F(u,v)|_{u=v} &= \partial_u \mathsf{F}(u,v)|_{u=v}\\ \mathsf{u}'(v)\partial_w F(v,w)|_{w=v} &= \partial_w \mathsf{F}(v,w)|_{w=v}\end{aligned} \qquad \forall v\in\Omega$$

is called c-consistent. An entropy flux which is Lax-consistent and c-consistent is called Lax-c-consistent.

If one considers

$$\begin{aligned}\mathsf{u}'(v)\partial_u F(u,v)|_{u=v} &= \mathsf{u}'(u)\partial_u\left(\frac{1}{2}[f(u)+f(v)]-\frac{1}{2\lambda}Q(u,v)[u-v]\right)|_{u=v}\\ &= \mathsf{u}'(u)\left(\frac{1}{2}f'(u)-\frac{1}{2\lambda}Q(u,u)\right)\\ &= \frac{1}{2}\mathsf{f}'(u)-\frac{1}{2\lambda}Q(u,u)\mathsf{u}'(u)\end{aligned}$$

and compare this with the derivation of the proposed numerical entropy flux with respect to u, i.e.

$$\begin{aligned}\partial_u \mathsf{F}(u,v)|_{u=v} &= \partial_u \left(\frac{1}{2}[f(u)+f(v)] - \frac{1}{2\lambda}\mathsf{Q}(u,v)[\mathsf{u}(u)-\mathsf{u}(v)]\right)|_{u=v} \\ &= \frac{1}{2}\mathsf{f}'(u) - \frac{1}{2\lambda}\mathsf{Q}(u,u)\mathsf{u}'(u)\end{aligned}$$

one sees clearly, that the this means nothing else than

$$Q(u,u) = \mathsf{Q}(u,u). \tag{2.35}$$

Remark 2.37
Tadmor [107] developed a whole theory for three-point schemes of purely second-order accuracy in space which are entropy stable according to some numerical entropy fluxes. However, Sonar [102] showed that these schemes fail to be entropy stable if a Lax-c-consistent numerical entropy flux is used.

(2.35) leads us to a much simpler form for the numerical entropy flux:

Definition 2.38
The numerical entropy flux

$$\mathsf{F}(u,v) := \frac{1}{2}[\mathsf{f}(u)+\mathsf{f}(v)] - \frac{1}{2\lambda}Q(u,v)[\mathsf{u}(u)-\mathsf{u}(v)] \tag{2.36}$$

is called the corresponding numerical entropy flux *to a numerical scheme of the form*

$$F(u,v) := \frac{1}{2}[f(u)+f(v)] - \frac{1}{2\lambda}Q(u,v)[u-v].$$

This means nothing else than $\mathsf{Q}(u,v) = Q(u,v)$.

The above considerations have shown that this choice is reasonable and is still Lax-c-consistent proving the following:

Lemma 2.39
The numerical entropy flux of the form (2.36) is Lax-c-consistent.

Proof Lax-consistency was shown above. The c-consistency is trivial since the entropy flux is chosen such that the dissipation coefficients are equal for the numerical flux and the numerical entropy flux. ∎

Note that in [15] Coquel and LeFloch use this kind of numerical entropy flux to derive sharp entropy inequalities for the modified Lax-Friedrichs scheme.

In the remainder of the section we show that the choice of the numerical entropy flux is consistent with the choice given by Crandall and Majda [20]. For a three-point scheme this is

$$\begin{aligned}\mathsf{F}^{CM}_{i+1/2} &= F(\max(U_{i+1},k),\max(U_i,k)) \\ &\quad -F(\min(U_{i+1},k),\min(U_i,k)), \quad k\in\mathbb{R}.\end{aligned} \tag{2.37}$$

Theorem 2.40
Assuming bounded data U , i.e.

$$|U_{i+1} - U_i| \leq ch \quad \forall i, \tag{2.38}$$

and $c \in \mathbb{R}^+$ sufficiently small, then the numerical entropy flux of the form

$$\mathsf{F}_{i+1/2} \;=\; \frac{1}{2}(\mathsf{f}_{i+1} + \mathsf{f}_i) - \frac{1}{2\lambda} Q_{i+1/2}(\mathsf{u}_{i+1} - \mathsf{u}_i) \tag{2.39}$$

with dissipation coefficient Q is a second-order accurate approximation to the Crandall-Majda entropy flux (2.37).

Proof Considering the Kruzkov entropy-pairs (1.25) with

$$\begin{aligned} \max(U_{i+1/i}, k) &=: (U_{i+1/i} \top k), \\ \min(U_{i+1/i}, k) &=: (U_{i+1/i} \bot k), \end{aligned}$$

we rewrite the entropy flux (2.37) as

$$\begin{aligned}
\mathsf{F}_{i+1/2} &= F(U_{i+1}\top k, U_i \top k) - F(U_{i+1}\bot k, U_i \bot k) \\
&= \frac{1}{2}[f(U_{i+1}\top k) + f(U_i \top k)] \\
&\quad -\frac{1}{2\lambda} Q(U_{i+1}\top k, U_i \top k)(U_{i+1}\top k - U_i \top k) \\
&\quad -\frac{1}{2}[f(U_{i+1}\bot k) + f(U_i \bot k)] \\
&\quad +\frac{1}{2\lambda} Q(U_{i+1}\bot k, U_i \bot k)(U_{i+1}\bot k - U_i \bot k) \\
&= \frac{1}{2}[f(U_{i+1}\top k) - f(U_{i+1}\bot k)] + \frac{1}{2}[f(U_i \top k) - f(U_i \bot k)] \\
&\quad -\frac{1}{2\lambda} Q(U_{i+1}\top k, U_i \top k)(U_{i+1}\top k - U_i \top k) \\
&\quad +\frac{1}{2\lambda} Q(U_{i+1}\bot k, U_i \bot k)(U_{i+1}\bot k - U_i \bot k) \\
&= \frac{1}{2}\mathrm{sign}(U_{i+1} - k)(f(U_{i+1}) - f(k)) + \frac{1}{2}\mathrm{sign}(U_i - k)(f(U_i) - f(k)) \\
&\quad -\frac{1}{2\lambda} Q(U_{i+1}\top k, U_i \top k)(U_{i+1}\top k - U_i \top k) \\
&\quad +\frac{1}{2\lambda} Q(U_{i+1}\bot k, U_i \bot k)(U_{i+1}\bot k - U_i \bot k) \\
&= \frac{1}{2}[\mathsf{f}(U_{i+1}) + \mathsf{f}(U_i)] \\
&\quad -\frac{1}{2\lambda} Q(U_{i+1}\top k, U_i \top k)(U_{i+1}\top k - U_i \top k) \\
&\quad +\frac{1}{2\lambda} Q(U_{i+1}\bot k, U_i \bot k)(U_{i+1}\bot k - U_i \bot k)
\end{aligned}$$

Hence, we have to prove that

$$\begin{aligned} Q(U_{i+1}\top k, U_i \top k)(U_{i+1}\top k - U_i \top k) - Q(U_{i+1}\bot k, U_i \bot k)(U_{i+1}\bot k - U_i \bot k) \\ = Q(U_{i+1}, U_i)(U_{i+1} - U_i) + O(h^2) \end{aligned}$$

holds. We analyse the following cases:

i) $k \geq U_{i+1}, U_i$:
$\Longrightarrow\ (U_{i/i+1} \top k) = k,\ (U_{i/i+1} \bot k) = U_{i/i+1}$:

$$\begin{aligned}
&-Q(k,k)[k-k] + Q(U_{i+1},U_i)[U_{i+1}-U_i] \\
&\quad = Q(U_{i+1},U_i)[U_{i+1}-U_i+k-k] \\
&\quad = Q(U_{i+1},U_i)[(U_{i+1}-k)-(U_i-k)] \\
&\quad = Q(U_{i+1},U_i)[\mathrm{sign}(U_{i+1}-k)|U_{i+1}-k| - \mathrm{sign}(U_i-k)|U_i-k|] \\
&\quad = -Q(U_{i+1},U_i)[\mathsf{u}_{i+1}-\mathsf{u}_i]
\end{aligned}$$

ii) $k \leq U_{i+1}, U_i$:
$\Longrightarrow\ (U_{i/i+1} \top k) = U_{i/i+1},\ (U_{i/i+1} \bot k) = k$:

$$\begin{aligned}
&-Q(U_{i+1},U_i)[U_{i+1}-U_i] + Q(k,k)[k-k] \\
&\quad = -Q(U_{i+1},U_i)[U_{i+1}-U_i+k-k] \\
&\quad = -Q(U_{i+1},U_i)[(U_{i+1}-k)-(U_i-k)] \\
&\quad = -Q(U_{i+1},U_i)[\mathrm{sign}(U_{i+1}-k)|U_{i+1}-k| - \mathrm{sign}(U_i-k)|U_i-k|] \\
&\quad = -Q(U_{i+1},U_i)[\mathsf{u}_{i+1}-\mathsf{u}_i]
\end{aligned}$$

iii) $U_{i+1} > k > U_i$:
$\Longrightarrow\ (U_{i+1} \top k) = U_{i+1},\ (U_i \top k) = k,\ (U_{i+1} \bot k) = k,\ (U_i \bot k) = U_i$:

$$\begin{aligned}
&Q(U_{i+1},k)[U_{i+1}-k] - Q(k,U_i)[k-U_i] \\
&\quad = Q(U_{i+1},k)[U_{i+1}-k] + Q(k,U_i)[U_i-k] \\
&\quad = Q(U_{i+1},k)|U_{i+1}-k| - Q(k,U_i)|U_i-k| \\
&\quad = Q(U_{i+1},k)\mathsf{u}_{i+1} - Q(k,U_i)\mathsf{u}_i
\end{aligned}$$

iv) $U_{i+1} < k < U_i$:
$\Longrightarrow\ (U_{i+1} \top k) = k,\ (U_i \top k) = U_i,\ (U_{i+1} \bot k) = U_{i+1},\ (U_i \bot k) = k$:

$$\begin{aligned}
&Q(k,U_i)[k-U_i] - Q(U_{i+1},k)[U_{i+1}-k] \\
&\quad = -Q(k,U_i)[U_i-k] - Q(U_{i+1},k)[U_{i+1}-k] \\
&\quad = -Q(k,U_i)|U_i-k| + Q(U_{i+1},k)|U_{i+1}-k| \\
&\quad = -Q(k,U_i)\mathsf{u}_i + Q(U_{i+1},k)\mathsf{u}_{i+1}
\end{aligned}$$

With the Taylor-expansions and the setting $k = \alpha U_i + (1-\alpha)U_{i+1}, \quad \alpha \in (0,1)$, i.e.

$$\begin{aligned}
Q(U_{i+1},k) &= Q(U_{i+1},U_i) + \partial_{U_i} Q(U_{i+1},U_i)(k-U_i) + O(h^2) \\
&= Q(U_{i+1},U_i) + \partial_{U_i} Q(U_{i+1},U_i)(1-\alpha)(U_{i+1}-U_i) + O(h^2)
\end{aligned}$$

$$\begin{aligned}
Q(k,U_i) &= Q(U_{i+1},U_i) - \partial_{U_{i+1}} Q(U_{i+1},U_i)(U_{i+1}-k) + O(h^2) \\
&= Q(U_{i+1},U_i) - \partial_{U_{i+1}} Q(U_{i+1},U_i)\alpha(U_{i+1}-U_i) + O(h^2),
\end{aligned}$$

we derive for the last case

$$\begin{aligned}
&Q(k,U_i)[k-U_i]-Q(U_{i+1},k)[U_{i+1}-k]\\
&= -Q(k,U_i)\mathsf{u}_i+Q(U_{i+1},k)\mathsf{u}_{i+1}\\
&= Q(U_{i+1},U_i)[\mathsf{u}_{i+1}-\mathsf{u}_i]\\
&\quad +\partial_{U_i}Q(1-\alpha)(U_{i+1}-U_i)\mathsf{u}_{i+1}-\partial_{U_{i+1}}Q\alpha(U_{i+1}-U_i)\mathsf{u}_i+O(h^3)\\
&= Q(U_{i+1},U_i)[\mathsf{u}_{i+1}-\mathsf{u}_i]\\
&\quad +\partial_{U_i}Q(1-\alpha)(U_{i+1}-U_i)\mathrm{sign}(U_{i+1}-k)(U_{i+1}-\alpha U_i-(1-\alpha)U_{i+1})\\
&\quad -\partial_{U_{i+1}}Q\alpha(U_{i+1}-U_i)\mathrm{sign}(U_i-k)(U_i-\alpha U_i-(1-\alpha)U_{i+1})+O(h^3)\\
&= Q(U_{i+1},U_i)[\mathsf{u}_{i+1}-\mathsf{u}_i]\\
&\quad -\left[\partial_{U_i}Q\alpha(1-\alpha)(U_{i+1}-U_i)^2+\partial_{U_{i+1}}Q\alpha(1-\alpha)(U_{i+1}-U_i)^2\right]+O(h^3)\\
&= Q(U_{i+1},U_i)[\mathsf{u}_{i+1}-\mathsf{u}_i]\\
&\quad -(hu'_{i+1/2})^2\alpha(1-\alpha)\left[\partial_{U_i}Q-\partial_{U_{i+1}}Q\right]+O(h^3)
\end{aligned}$$

Since the function $f(\alpha)=\alpha(1-\alpha)$ achieves its maximum at $\alpha=1/2$ with $f(1/2)=1/4$ and we assume data of the form (2.38) one gets

$$\begin{aligned}
|-Q(U_{i+1},k)[U_{i+1}-k]+Q(k,U_i)[k-U_i]|-|Q(U_{i+1},U_i)[\mathsf{u}_{i+1}-\mathsf{u}_i]|&\\
\leq \frac{1}{4}h^2c^2|(\partial_{U_{i+1}}-\partial_{U_i})Q(U_{i+1},U_i)|+O(h^3).&
\end{aligned}$$

which proves the Theorem.

■

Remark 2.41
Numerical entropy fluxes of the form (2.36) were already used in [37] for the computation of numerical entropy inequalities. There they served as an entropy indicator to distinguish between regions with discontinuities and smooth ones. The results were quite satisfying.

Discrete cell entropy inequality

Following Merriam [80, 81], we start from the semi-discrete form of the entropy inequality (1.24), i.e.

$$\tfrac{d}{dt}\,\mathsf{u}(u)_i+\tfrac{1}{\Delta x_1}[\mathsf{F}_{i+1/2,j}-\mathsf{F}_{i-1/2,j}] \;\leq\; 0. \tag{2.40}$$

Now we are interested in computing a discrete cell entropy inequality for each cell, in order to decouple the discrete entropy inequality (2.34) applied to a three-point scheme. Since we assume piecewise constant data U_i over the cell C_i, with

$$\mathsf{u}_i \;=\; \mathsf{u}([\overline{u}]_i) \;=\; \mathsf{u}(U_i),$$

with

$$\begin{aligned} U_i &= [\overline{u}]_i \\ &= \frac{1}{h}\int\limits_{C_i} u(\xi, t^n)\, d\xi, \end{aligned}$$

the chain rule is valid:

$$\partial_t \mathsf{u}(U_i) \;=\; \mathsf{u}'\partial_t(U_i) \;=\; \mathsf{u}'(u_t)_i. \tag{2.41}$$

Substituting this into the semi-discrete cell inequality (2.40) using the numerical approximation for the time derivative, i.e. the difference of the numerical fluxes yields the discrete cell entropy inequality for the cell C_i:

$$E_i \;:=\; -\mathsf{u}_i'[F_{i+1/2} - F_{i-1/2}] + [\mathsf{F}_{i+1/2} - \mathsf{F}_{i-1/2}] \le 0. \tag{2.42}$$

Such equations using $U_{i+1/2}$ and $U_{i-1/2}$ as interpolants at cell boundaries were extensively studied by Merriam [81]. He splits (2.42) into two parts $E_{i+1/2}, E_{i-1/2}$ related to the corresponding cell boundaries, i.e.

$$\begin{aligned} E_i &= E_{i+1/2} + E_{i-1/2} \\ &= \underbrace{[-\mathsf{u}_i' F_{i+1/2} + \mathsf{F}_{i+1/2}]}_{E_{i+1/2}} + \underbrace{[\mathsf{u}_i' F_{i-1/2} - \mathsf{F}_{i-1/2}]}_{E_{i-1/2}} \le 0. \end{aligned}$$

In order to satisfy the discrete entropy inequality, i.e. $E_i \le 0$, one can follow the more restrictive approach to match the cell boundary inequalities

$$\begin{aligned} E_{i+1/2} &\le 0, \\ E_{i-1/2} &\le 0. \end{aligned} \tag{2.43}$$

This approach is easier to handle since it decouples the discrete cell entropy inequality (2.42) into two inequalities for each cell boundary.

With the proposed corresponding numerical entropy flux (2.36) above, we are able to write the cell boundary entropy inequalities (2.43) as

$$\begin{aligned} E_{i+1/2} &= -\mathsf{u}_i'\left[\frac{1}{2}(f_{i+1} + f_i) - \frac{1}{2\lambda}Q_{i+1/2}(U_{i+1} - U_i)\right] \\ &\quad + \frac{1}{2}(\mathsf{f}_{i+1} - \mathsf{f}_i) - \frac{1}{2\lambda}Q_{i+1/2}(\mathsf{u}_{i+1} - \mathsf{u}_i) \end{aligned}$$

resp.

$$\begin{aligned} E_{i-1/2} &= \mathsf{u}_i'\left[\frac{1}{2}(f_i + f_{i-1}) - \frac{1}{2\lambda}Q_{i-1/2}(U_i - U_{i-1})\right] \\ &\quad - \frac{1}{2}(\mathsf{f}_i - \mathsf{f}_{i-1}) - \frac{1}{2\lambda}Q_{i-1/2}(\mathsf{u}_i - \mathsf{u}_{i-1}). \end{aligned}$$

The interesting result and the reward for the special choice of the numerical entropy flux given above is the following

Lemma 2.42
A scheme of the form (2.7) with corresponding numerical entropy flux (2.36) satisfies the discrete cell entropy inequality (2.42) if one chooses the numerical dissipation coefficient $Q_{i+1/2}$ to be

$$Q^E_{i+1/2} = Q^E(U_{i+1}, U_i) \quad := \quad \max(Q^L_{i+1/2}, Q^R_{i+1/2}),$$

with

$$Q^R_{i+1/2} \quad \geq \quad -\lambda \frac{[(\mathsf{f}_{i+1} - \mathsf{f}_i) - \mathsf{u}'_{i+1}(f_{i+1} - f_i)]}{[(\mathsf{u}_{i+1} - \mathsf{u}_i) - \mathsf{u}'_{i+1}(U_{i+1} - U_i)]}, \tag{2.44}$$

$$Q^L_{i+1/2} \quad \geq \quad \lambda \frac{[(\mathsf{f}_{i+1} - \mathsf{f}_i) - \mathsf{u}'_{i}(f_{i+1} - f_i)]}{[(\mathsf{u}_{i+1} - \mathsf{u}_i) - \mathsf{u}'_{i}(U_{i+1} - U_i)]}. \tag{2.45}$$

Proof We start from the restrictive demand (2.43) and examine the inequality for $E_{i+1/2} \leq 0$, i.e.

$$\begin{aligned} E_{i+1/2} &= [-\mathsf{u}'_i F_{i+1/2} + \mathsf{F}_{i+1/2}] \\ &= -\mathsf{u}'_i \left[\frac{1}{2}(f_{i+1} + f_i) - \frac{1}{2\lambda} Q_{i+1/2}(U_{i+1} - U_i) \right] \\ &\quad + \frac{1}{2}(\mathsf{f}_{i+1} - \mathsf{f}_i) - \frac{1}{2\lambda} Q_{i+1/2}(\mathsf{u}_{i+1} - \mathsf{u}_i) \leq 0. \end{aligned}$$

Rearranging this inequality in order to derive an inequality for the numerical dissipation coefficient $Q_{i+1/2}$ yields (2.44).

If we do the same for the cell inequality of the neighbouring cell C_{i+1} and consider E^L_{i+1}, we derive (2.45). The maximum satisfies both cell boundary entropy inequalities concerning the boundary $C_{i+1/2}$. Doing the same for $C_{i-1/2}$ reveals the desired result. ∎

Remark 2.43
Remember that we have used the quite restrictive form (2.43). Thus, there might be hope to follow the path where we admit the violation of (2.43) and require only the discrete cell inequality (2.42). Obviously more degrees of freedom result in a much more complex analysis since this leads to a coupled system for the dissipation coefficients.

One can think of the use of linear programming to minimise the entropy over all cells under the requirement that (2.42) is satisfied in each cell. From the theoretical point of view this may be an interesting scheme even if in practice this will lead to a very slow algorithm.

Corollary 2.44
Up to second order the numerical dissipation coefficient $Q^E_{i+1/2}$ coincides with the dissipation coefficient of the Roe scheme [92] or Murman-Courant-Isaacson-Rees scheme [84, 18], i.e.

$$Q^E_{i+/12} \quad = \quad \lambda \left| f'(U_{i+1/2}) \right| + O(h^2)$$

Proof First (2.45) is considered. If we use Taylor-series expansion the expression reads

$$\lambda \left[\frac{h[\mathsf{f}'(U_{i+1/2})u'_{i+1/2} - \mathsf{u}'(U_i)f'(U_{i+1/2})u'_{i+1/2}] + O(h^2)}{h[\mathsf{u}'(U_{i+1/2})u'_{i+1/2} - \mathsf{u}'(U_i)u'_{i+1/2}] + O(h^2)} \right].$$

Since the entropy flux obeys the compatibility relation (1.22), one gets

$$\lambda \left[\frac{h[\mathsf{u}'(U_{i+1/2}) - \mathsf{u}'(U_i)] f'(U_{i+1/2}) u'_{i+1/2} + O(h^2)}{h[\mathsf{u}'(U_{i+1/2}) - \mathsf{u}'(U_i)] u'_{i+1/2} + O(h^2)} \right].$$

Thus, we derive

$$\lambda \frac{f'(U_{i+1/2}) + O(h^2)}{1 + O(h^2)} \quad = \quad \lambda f'(U_{i+1/2}) + O(h^2).$$

On the other hand, if we examine (2.44) by using a Taylor-series expansion the expression reads as

$$-\lambda \left[\frac{h[\mathsf{f}'(U_{i+1/2}) u'_{i+1/2} - \mathsf{u}'(U_{i+1}) f'(U_{i+1/2}) u'_{i+1/2}] + O(h^2)}{h[\mathsf{u}'(U_{i+1/2}) u'_{i+1/2} - \mathsf{u}'(U_{i+1}) u'_{i+1/2}] + O(h^2)} \right].$$

Similar to the case above, we derive

$$-\lambda \left[\frac{h[\mathsf{u}'(U_{i+1/2}) - \mathsf{u}'(U_{i+1})] f'(U_{i+1/2}) u'_{i+1/2} + O(h^2)}{h[\mathsf{u}'(U_{i+1/2}) - \mathsf{u}'(U_{i+1})] u'_{i+1/2} + O(h^2)} \right],$$

and end up with

$$-\lambda \frac{f'(U_{i+1/2}) + O(h^2)}{1 + O(h^2)} \quad = \quad -\lambda f'(U_{i+1/2}) + O(h^2).$$

Since the cell boundary entropy inequality satisfying the derived dissipation coefficient is defined as the maximum over both expressions (2.44),(2.45), we get

$$Q^E_{i+1/2} \quad = \quad \lambda |f'_{i+1/2}| + O(h^2)$$

which is a second order approximation of

$$Q^{Roe}_{i+1/2} \quad = \quad \lambda \left| \frac{f(U_{i+1}) - f(U_i)}{U_{i+1} - U_i} \right|.$$

■

If we look in detail at the expressions in (2.44),(2.45) there are some interesting consequences. Emanating from (2.45), i.e.

$$\lambda \frac{[(\mathsf{f}_{i+1} - \mathsf{f}_i) - \mathsf{u}'_i (f_{i+1} - f_i)]}{[(\mathsf{u}_{i+1} - \mathsf{u}_i) - \mathsf{u}'_i (U_{i+1} - U_i)]},$$

and examining the denominator first, the convexity of $\mathsf{u}(u)$ immediately gives

$$(\mathsf{u}_{i+1} - \mathsf{u}_i) - \mathsf{u}'_i (U_{i+1} - U_i) \quad \geq \quad 0. \tag{2.46}$$

Using the Rankine-Hugoniot condition for the entropy inequality, i.e.

$$s[\mathsf{u}(u)] \geq [\mathsf{f}(u)]$$

with $s = (f_r - f_l)/(u_r - u_l)$ the entropy condition reads as

$$[\mathsf{f}(u_r) - \mathsf{f}(u_l)] - \left(\frac{\mathsf{u}(u_r) - \mathsf{u}(u_l)}{u_r - u_l}\right)[f(u_r) - f(u_l)] \quad \leq \quad 0. \tag{2.47}$$

Inserting the Taylor-approximation for $\mathsf{u}(U_{i+1})$, i.e. the numerator of (2.45) with $u_r = U_{i+1}$, $u_l = U_i$, one gets

$$[\mathsf{f}(U_{i+1}) - \mathsf{f}(U_i)] - \mathsf{u}_i'[f(U_{i+1}) - f(U_i)] \quad \leq \quad 0.$$

We assume

$$\mathsf{u}_i'(f_{i+1} - f_i) \quad \geq \quad \frac{\mathsf{u}_{i+1} - \mathsf{u}_i}{U_{i+1} - U_i}(f_{i+1} - f_i).$$

This inequality depends on the signs of $f_{i+1} - f_i$ and $U_{i+1} - U_i$ and holds for $\operatorname{sign}(\Delta f_{i+1/2}), \operatorname{sign}(\Delta U_{i+1/2})$ both positive or both negative simultaneously. With the convexity requirement (2.46) it follows immediately by switching signs and adding the difference of the entropy fluxes that

$$\begin{aligned}&[\mathsf{f}(U_{i+1}) - \mathsf{f}(U_i)] - \mathsf{u}_i'[f(U_{i+1}) - f(U_i)] \\ &\qquad \leq [\mathsf{f}(U_{i+1}) - \mathsf{f}(U_i)] - \left(\frac{\mathsf{u}(U_{i+1}) - \mathsf{u}(U_i)}{U_{i+1} - U_i}\right)[f(U_{i+1}) - f(U_i)].\end{aligned}$$

Since we have to assure the discrete Rankine-Hugoniot condition (2.47) this leads to

$$[\mathsf{f}(U_{i+1}) - \mathsf{f}(U_i)] - \mathsf{u}_i'[f(U_{i+1}) - f(U_i)] \quad \leq \quad 0.$$

We derive that the dissipation coefficient $Q^L_{i+1/2}$, (2.45) is nonpositive, since the denominator in (2.46) was nonnegative.

Considering the expression (2.44), the convexity requirement for the entropy function $\mathsf{u}(u)$ is

$$(\mathsf{u}_{i+1} - \mathsf{u}_i) - \mathsf{u}_{i+1}'(U_{i+1} - U_i) \leq 0.$$

To assure that

$$\begin{aligned}&-[\mathsf{f}(U_{i+1}) - \mathsf{f}(U_i)] + \mathsf{u}_{i+1}'[f(U_{i+1}) - f(U_i)] \\ &\qquad \geq -[\mathsf{f}(U_{i+1}) - \mathsf{f}(U_i)] + \left(\frac{\mathsf{u}(U_{i+1}) - \mathsf{u}(U_i)}{U_{i+1} - U_i}\right)[f(U_{i+1}) - f(U_i)] \geq 0\end{aligned}$$

we start with the first inequality. This holds if one guarantees

$$\mathsf{u}_{i+1}'(f_{i+1} - f_i) \quad \geq \quad \frac{\mathsf{u}_{i+1} - \mathsf{u}_i}{U_{i+1} - U_i}(f_{i+1} - f_i)$$

which is again true for $f_{i+1} - f_i$ and $U_{i+1} - U_i$ having the same sign and we have

$$-[\mathsf{f}(U_{i+1}) - \mathsf{f}(U_i)] + \mathsf{u}_i'[f(U_{i+1}) - f(U_i)] \ \geq \ 0.$$

This leads to the fact, that (2.44) may be nonpositive. These considerations show that it is possible that even the maximum of (2.44),(2.45) might be negative.

This fact can also be illustrated by an instructive example. First we rewrite the dissipation coefficients as

$$\begin{aligned} Q^R_{i+1/2} &= \lambda \frac{[(\mathsf{f}_{i+1}-\mathsf{f}_i)-\mathsf{u}'_{i+1}(f_{i+1}-f_i)]}{[\mathsf{u}'_i(U_{i+1}-U_i)-(\mathsf{u}_{i+1}-\mathsf{u}_i)]}, \\ Q^L_{i+1/2} &= \lambda \frac{[(\mathsf{f}_{i+1}-\mathsf{f}_i)-\mathsf{u}'_i(f_{i+1}-f_i)]}{[(\mathsf{u}_{i+1}-\mathsf{u}_i)-\mathsf{u}'_i(U_{i+1}-U_i)]}. \end{aligned}$$

By the convexity argument both denominators are positive and we focus on the nominator. If the entropy flux $\mathsf{f}(u)$ has a minimum at $U_{i+1/2}$ i.e. $\mathsf{f}'_{i+1/2} = \mathsf{u}'_{i+1/2} f'_{i+1/2} = 0$ we might have

$$\begin{aligned} \mathsf{u}'_{i+1}(f_{i+1}-f_i) &> (\mathsf{f}_{i+1}-\mathsf{f}_i) = 0, \\ \mathsf{u}'_i(f_{i+1}-f_i) &> (\mathsf{f}_{i+1}-\mathsf{f}_i) = 0, \end{aligned}$$

and both numerators are negative and so are the dissipation coefficients. To remedy this we add a positive constant δ which avoids that the dissipation coefficient is negative or disappears:

$$Q_{i+1/2} = \max(\delta, Q^L, Q^R).$$

Consequently, the above consideration proves the following

Theorem 2.45
The dissipation coefficient of the form

$$Q^{CE}_{i+1/2} = \lambda \max(\delta, Q^L, Q^R), \qquad 0 < \delta \ll 1, \tag{2.48}$$

derived from the cell entropy inequalities (2.44), (2.45) is consistent with the Roe scheme with entropy fix.

Remark 2.46
This correction is by no means surprising. Since we have seen that the limit of the cell entropy coefficients (2.44),(2.45) gives the dissipation coefficient of the Roe-scheme. This scheme is known to produce entropy-violating solutions and needs an entropy fix given by Harten and Hyman [45]. The fix is of the same kind as in our case. If the numerical dissipation coefficient gets to small, an ε takes the role of the dissipation coefficient, i.e.

$$Q^{Roe}_{i+1/2} = \begin{cases} \lambda \left| \frac{\Delta_{i+1/2} f}{\Delta_{i+1/2} u} \right| & \text{for } \lambda \left| \frac{\Delta_{i+1/2} f}{\Delta_{i+1/2} u} \right| > \varepsilon, \\ \varepsilon & \lambda \left| \frac{\Delta_{i+1/2} f}{\Delta_{i+1/2} u} \right| \le \varepsilon, \end{cases} \qquad \varepsilon \ll 1.$$

Lemma 2.47
The dissipation coefficient (2.48) can be written as

$$Q^{CE}_{i+1/2} = \max_{u^\star \in [U_i, U_{i+1}]} \left[\lambda \left(\frac{f_{i+1} - f(u^\star)}{U_{i+1} - u^\star} \right), -\lambda \left(\frac{f(u^\star) - f_i}{u^\star - U_i} \right), \delta \right].$$

Proof As we know from the result of Kruzkov [62] it is sufficient to consider the family of entropy pairs

$$\mathsf{u}_i = |U_i - k|, \quad \mathsf{f}_i = \text{sign}(U_i - k)(f_i - f(k)), \quad \forall k \in \mathbb{R}.$$

If we insert this into the expressions (2.45) we derive

$$\begin{aligned} Q^L_{i+1/2} &= \lambda \left[\frac{\mathsf{f}_{i+1} - \mathsf{f}_i - \text{sign}(U_i - k)(f_{i+1} - f_i)}{|U_{i+1} - k| - |U_i - k| - \text{sign}(U_i - k)(U_{i+1} - U_i)} \right] \\ &= \lambda \left[\frac{\mathsf{f}_{i+1} - \mathsf{f}_i - \text{sign}(U_i - k)(f_{i+1} - f(k) + f(k) - f_i)}{|U_{i+1} - k| - |U_i - k| - \text{sign}(U_i - k)(U_{i+1} - k + k - U_i)} \right] \\ &= \lambda \left[\frac{\mathsf{f}_{i+1} - \mathsf{f}_i + \mathsf{f}_i - \text{sign}(U_i - k)[f_{i+1} - f(k)]}{|U_{i+1} - k| - |U_i - k| + |U_i - k| - \text{sign}(U_i - k)(U_{i+1} - k)} \right] \\ &= \lambda \left[\frac{[\text{sign}(U_{i+1} - k) - \text{sign}(U_i - k)](f_{i+1} - f(k))}{[\text{sign}(U_{i+1} - k) - \text{sign}(U_i - k)](U_{i+1} - k)} \right] \end{aligned}$$

As we have seen in the proof of Theorem 2.40 the interesting case is $k \in (U_i, U_{i+1})$ for $U_i < U_{i+1}$ resp. $k \in (U_{i+1}, U_i)$ for $U_i > U_{i+1}$. This gives the desired form

$$\lambda \left[\frac{f_{i+1} - f(u^\star)}{U_{i+1} - u^\star} \right], \quad u^\star \in (U_i, U_{i+1}).$$

The same computation for $Q^R_{i+1/2}$ gives the analogous result and finally the form given above. ∎

Corollary 2.48
A three-point scheme with numerical dissipation coefficient (2.48), i.e.

$$Q^{CE}_{i+1/2} = \max(\delta, Q^L, Q^R), \qquad 0 < \delta \ll 1, \tag{2.49}$$

is an E-scheme.

Proof Starting from the definition for the flux of an E-scheme one has

$$\text{sign}(U_{i+1} - U_i)[F_{i+1/2} - f(u)] \leq 0, \quad \forall u \in [U_i, U_{i+1}].$$

Using the numerical flux function

$$F^{CE}_{i+1/2} = \frac{1}{2}(f_{i+1} + f_i) - \frac{1}{2\lambda} Q^{CE}_{i+1/2}(U_{i+1} - U_i)$$

one gets

$$
\begin{aligned}
&\operatorname{sign}(U_{i+1}-U_i)[F^{CE}_{i+1/2}-f(u)]\\
&\quad= \operatorname{sign}(U_{i+1}-U_i)\left[\frac{1}{2}(f_{i+1}+f_i)-\frac{1}{2\lambda}Q^{CE}_{i+1/2}(U_{i+1}-U_i)-f(u)\right]\\
&\quad= \operatorname{sign}(U_{i+1}-U_i)\left[\frac{1}{2}(f_{i+1}-f(u))+\frac{1}{2}(f_i-f(u))-\frac{1}{2\lambda}Q^{CE}_{i+1/2}(U_{i+1}-U_i)\right]\\
&\quad= \operatorname{sign}(U_{i+1}-U_i)\left[\frac{1}{2}(f_{i+1}-f(u))-\frac{1}{2}(f(u)-f_i)-\frac{1}{2\lambda}Q^{CE}_{i+1/2}(U_{i+1}-U_i)\right]\\
&\quad= \operatorname{sign}(U_{i+1}-U_i)\left[(U_{i+1}-u)\left(\frac{1}{2}\frac{f_{i+1}-f(u)}{U_{i+1}-u}-\frac{1}{2\lambda}Q^{CE}_{i+1/2}\right)\right.\\
&\qquad\qquad\left.-(u-U_i)\left(\frac{1}{2}\frac{f(u)-f_i}{u-U_i}-\frac{1}{2\lambda}Q^{CE}_{i+1/2}\right)\right]\\
&\quad= \operatorname{sign}(U_{i+1}-U_i)\left[-(U_{i+1}-u)\underbrace{\left(\frac{1}{2\lambda}Q^{CE}_{i+1/2}-\frac{1}{2}\frac{f_{i+1}-f(u)}{U_{i+1}-u}\right)}_{\alpha}\right.\\
&\qquad\qquad\left.-(u-U_i)\underbrace{\left(\frac{1}{2\lambda}Q^{CE}_{i+1/2}-\frac{1}{2}\frac{f(u)-f_i}{u-U_i}\right)}_{\beta}\right].
\end{aligned}
$$

By Lemma 2.47, the dissipation coefficient $Q^{CE}_{i+1/2}$ majorises the terms on the right-hand side, i.e. $\alpha,\beta\geq 0$. Therefore, the terms in the square brackets each are nonnegative and we get

$$
\begin{aligned}
&\operatorname{sign}(U_{i+1}-U_i)[F^{CE}_{i+1/2}-f(u)]\\
&\quad= \operatorname{sign}(U_{i+1}-U_i)[-(U_{i+1}-u)\alpha-(u-U_i)\beta]\\
&\quad\leq -\operatorname{sign}(U_{i+1}-U_i)[(U_{i+1}-U_i)\max(\alpha,\beta)] = -|U_{i+1}-U_i|\max(\alpha,\beta)\leq 0,
\end{aligned}
$$

which is equal to the Definition 2.30 of an E-scheme. ∎

Adaptive smoothing methods are based on the idea of applying a process which itself depends on local properties of the image

J.Weickert ("Anisotropic Diffusion in Image Processing" [116])

3 Dissipation filters

As already mentioned in the previous chapter, a numerical discretisation of high-order needs a stabilising term called numerical diffusion. This term can be seen as a regularisation or filter term which erases the oscillations.

In the area of numerical approximation to solutions of conservation laws it became somehow fashionable to interpret algorithms as convection steps propagating the solution in the space-time domain by a high-order method, followed by a filter step. This second step should remove the oscillations occurring due to the nonlinear instabilities related to high-order methods at discontinuities.

This view at a numerical approximation of a conservation law is somehow reverse to the Flux-Corrected-Transport (FCT) approach by Boris and Book [8, 9]. There, one starts from a monotone solution and adds an antidiffusive flux in smooth areas to guarantee second-order accuracy. The filter approach is related to the artificial-viscosity or modified flux approach by Harten [50, 42, 43] but splits in two steps. Thus, one defines $\mathcal{C}$ as the high-order operator convecting the solution and $\mathcal{F}$ as the filter operator filtering high frequencies, i.e. oscillations in the vicinity of discontinuities. The numerical solution at time level $n+1$ may be written as

$$\begin{aligned} u^{n+1/2} &= \mathcal{C}(\Delta t, \Delta x, u^n)u^n \\ u^{n+1} &= \mathcal{F}(\Delta t, \Delta x, u^{n+1/2})u^{n+1/2}. \end{aligned}$$

Engquist, Lötstedt and Sjögreen [25] were the first in constructing a family of filter algorithms post-processing a numerical solution gained by the Lax-Wendroff scheme in order to remove over- and undershoots. The resulting filter was proven to be second-order and TVD in the doctoral thesis of Sjögreen [100].

Lafon and Osher [63] followed this approach by building a numerical approximation using central differences of arbitrary high-order accuracy as the convection algorithm and built the (or several) filter step(s) from Essentially Non Oscillatory (ENO)-schemes (see [48, 49, 44]). Since ENO-schemes are computationally very costly this approach possesses the advantage of using simple central differences which are fast as compared to ENO, and requires the whole ENO machinery only at regions producing spurious oscillations, i.e. where it is really necessary.

After these initial approaches a huge amount of papers were published concentrating on the

Theory of digital image processing	Computational Fluid Dynamics
– Digital image	– Numerical solution of a two dimensional problem
– Image window	– Stencil of a difference scheme
– Pixel	– Cell of a spatial grid
– Edge	– Strong discontinuity
– Edge blur	– Smearing of a strong discontinuity
– Image noise	– Parasitic oscillations of the numerical solution
– Edge detection	– Localisation of discontinuities
– Smoothing of the image intensity function	– Smoothing of the numerical solution

Table 3.1: Comparison of terms used in image processing and CFD

interpretation or creation of filter algorithms for conservation laws.

Interestingly, similar tasks were carried out in other contexts. In image processing, mainly in image restoration and image enhancement, one is looking for numerical algorithms which on the one hand remove small scale oscillations, originating from the picture and on the other hand enhance the edges and corners, which determine the image.

In CFD, the origin of strong fronts is a well known fact, e.g. shock waves in gas dynamics and magneto-hydrodynamics, flame fronts in combustion theory and atmospheric and oceanic fronts in geophysical fluid dynamics. Thus, it is remarkable, that in both areas related methods could be used to handle the occurrence of singularities. However, it is astonishing, that numerical methods from this area are rarely used for the numerical treatment of conservation laws.

Vorozhtsov and Yanenko [112] use isotropic edge detectors for shock fitting algorithms and the localisation of singularities. To interpret algorithms from image processing and image restoration, they give a little *vocabulary* to interpret some notions of digital image processing in terms used in Computational Fluid Dynamics (see Table 3.1). They present several tools for image processing and pattern recognition and provide a broad knowledge of discrete filters and the control of artificial viscosity. Nevertheless, all approaches are based on linear ideas.

In this thesis, our goal is to build bridges between both fields, image processing and numerical approximations of conservation laws, which are based on nonlinear methods, and try to modify them for our purpose. Interestingly, the opposite way has already been explored (see [93, 86, 87, 99]). As we will see, the concepts of TVD-, ENO- and shock adapted dissipation methods are incorporated in the field of image processing. This work is oriented in the opposite direction: to use ideas and knowledge from image processing to build nonlinear dissipation models to control the numerical viscosity in order to stabilise the numerical algorithms and enhance shock fronts.

In the following sections, we will start by taking a look at algorithms from image processing. Subsequently, we present some classical diffusion filters in numerical schemes for conservation

laws. If no other reference is quoted, we refer to the excellent books of Aubert and Kornprobst [6] and Weickert [116].

In this section we are mostly concerned with the practical behaviour of the algorithms. However, since solutions of diffusion equations can be parametrised by their diffusion time a whole theory is developed in image processing concerning the so called scale space. Unfortunately, this does not apply to numerical approximations. There, the main interest is focused on a special diffusion strength, which cannot be cast in a scale theory reaching from scale nil, i.e unfiltered solution to infinity, i.e. equilibrium. An excellent overview over the scale-space theory is given by Lindeberg [75].

3.1 Linear filters

Linear filters are the most simple kind of diffusion filters available. They apply the same amount of filtering or diffusion to every point (pixel) of the data. So we get a data independent blurring of the signal.

Gaussian smoothing

A widely used way to smooth a signal represented by a real-valued mapping $u \in L^1(\mathbb{R}^2)$ is convolution with a Gaussian kernel:

$$(G_\sigma * u)(\underline{x}) := \int_{\mathbb{R}^2} G_\sigma(\underline{x} - \underline{y}) u(\underline{y})\, d\underline{y}. \tag{3.1}$$

G_σ represents the two-dimensional Gaussian with width (standard deviation) $\sigma = \sqrt{2t} > 0$ which reads as

$$\begin{aligned} G_\sigma(\underline{x}) &:= \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\underline{x}|^2}{2\sigma^2}\right) \\ &= \frac{1}{4\pi t} \exp\left(-\frac{|\underline{x}|^2}{4t}\right) \\ &= G(t, \underline{x}). \end{aligned}$$

From the convolution Theorem it follows that the Fourier transform of the convolution is equal to the product of the Fourier transform of the convolution kernel and the function u, i.e.

$$\mathcal{F}[G_\sigma * u](\underline{\omega}) = \mathcal{F}[G_\sigma](\underline{\omega}) \cdot \mathcal{F}[u](\underline{\omega}),$$

with the Fourier transform defined by

$$\mathcal{F}[u](\underline{\omega}) := \int_{\mathbb{R}^2} u(\underline{x}) \exp(-i\langle \underline{\omega}, \underline{x} \rangle)\, d\underline{x}.$$

The interesting, but not astonishing, fact is that the Fourier transform of a Gaussian shaped function is again of Gaussian form:

$$\mathcal{F}[G_\sigma](\omega) = \exp\left(-\frac{(|\omega|\sigma)^2}{2}\right). \tag{3.2}$$

Thus, it follows that

$$\mathcal{F}[G_\sigma * u](\omega) = \exp\left(-\frac{(|\omega|\sigma)^2}{2}\right)\mathcal{F}[u](\omega),$$

i.e the convolution with a Gaussian is a low-pass filter that inhibits frequencies (oscillations in the space domain). This damping of high frequencies in the signal u in a monotone way can be viewed as a diffusion process.

Linear diffusion equations

It is easy to see that the convolution of a signal u with a Gaussian kernel G_σ is a smoothing process. Since G_σ is a mollifier, high frequencies are damped and the total variation of the signal u is reduced. If we look at the smoothed signal

$$u_\sigma(\underline{x}) = \int_{\mathbb{R}^2} G_\sigma(\underline{x}-\underline{y})u(\underline{y})\,d\underline{y},$$

from the theory for linear partial differential equations we have the following

Theorem 3.1
The solution of the linear heat equation

$$\begin{aligned}\partial_t u &= \Delta u,\\ u(0,\underline{x}) &= u_0(\underline{x})\end{aligned}$$

with bounded initial data $u_0(\underline{x}) \in C(\mathbb{R}^2)$ is given by

$$\begin{aligned}u(t,\underline{x}) &= u_{\sigma^n}(\underline{x})\\ &= \int_{\mathbb{R}^2} G_\sigma(\underline{x}-\underline{y})u(t,\underline{y})\,d\underline{y}.\end{aligned}$$

The proof can be found e.g. in the books by John [57] or Evans [28].

From this well known fact one immediately sees that linear filtering of a signal u by convolution is equivalent to solving the linear heat equation for the the initial data u_0. If we restrict ourselves for a moment to one space dimension and look for a suitable discrete approximation of the heat equation, we see that the finite difference formulation

$$\begin{aligned}u(t,x+h)-2u(t,x)+u(t,x-h) &= u(t,x)+hu'(t,x)+\tfrac{1}{2}h^2u''(t,x)+\tfrac{1}{3}h^3u'''(t,x)\\ &\quad -2u(t,x)\\ &\quad +u(t,x)-hu'(t,x)+\tfrac{1}{2}h^2u''(t,x)-\tfrac{1}{3}h^3u'''(t,x)+\mathcal{O}(h^4)\\ &= h^2u''(t,x)+\mathcal{O}(h^4),\end{aligned}$$

i.e.

$$\frac{u(t,x+h)-2u(t,x)+u(t,x-h)}{h^2} = u''(t,x)+\mathcal{O}(h^2),$$

is an approximation to u_{xx} of second order accuracy. Equipped with a forward Euler approximation in time one derives the finite difference formulation for the heat equation

$$\frac{u(t^{n+1},x_i)-u(t^n,x_i)}{\Delta t} = \frac{u(t,x+h)-2u(t,x)+u(t,x-h)}{h^2}.$$

A time-advancing scheme for the solution of the heat equation consequently reads as

$$U^i = U_i + \frac{\Delta t}{h^2}[U_{i+1}-2U_i+U_{i-1}], \tag{3.3}$$

which is the simplest discrete model for a low pass filter. If we consider the Lax-Friedrichs scheme (2.27) in the foregoing chapter, the numerical viscosity term models a linear diffusion process. Having in mind the derivation of the Lax-Friedrichs scheme and its modification by Tadmor [105], we have the following

Lemma 3.2
The modified Lax-Friedrichs scheme (2.31), i.e.

$$\begin{aligned} U^i &= U_i - \lambda[F^{mLF}_{i+1/2} - F^{mLF}_{i-1/2}] \\ &= U_i - \lambda\left[f_{i+1/2} - \frac{1}{2\lambda}Q^{mLF}_{i+1/2}(U_{i+1}-U_i) - f_{i-1/2} + \frac{1}{2\lambda}Q^{mLF}_{i-1/2}(U_i-U_{i-1})\right] \\ &= U_i - \lambda[f_{i+1/2}-f_{i-1/2}] + \frac{1}{2}\left[Q^{mLF}_{i+1/2}(U_{i+1}-U_i) - Q^{mLF}_{i-1/2}(U_i-U_{i-1})\right] \\ &= U_i - \lambda[f_{i+1/2}-f_{i-1/2}] + \frac{1}{4}\left[U_{i+1}-2U_i-U_{i-1}\right] \end{aligned} \tag{3.4}$$

i) possesses a linear dissipation model,

ii) is the scheme with the highest tolerable numerical diffusion in order to get an entropy satisfying scheme.

Proof

i) This is easily seen from (3.4), which is a second-order accurate approximation to the desired conservation law $\partial_t u + \partial_x f(u) = 0$ augmented with the discretised form of the heat equation (3.3), weighted by the term $h^2/(4\Delta t)$.

ii) This is proved by Tadmor [106] as already mentioned in the second paragraph, since in Theorem 2.33 the modified Lax-Friedrichs scheme is noticed as the upper bound for an entropy-satisfying three-point scheme.

■

Since both schemes differ only by a factor of two for the numerical dissipation coefficient, this explains sufficiently the robustness and the dissipative behaviour of these algorithms.

The dissipation models of these schemes represent an integrated low-pass filter which avoids the tendency of the unstable second-order approximation of the flux function f to produce

spurious oscillations for non-smooth initial data. However, since this is a quite dissipative (and linear) filter it can not distinguish between areas in need of different smoothing strength, it tends to damp even structures like discontinuities.

Since we have already seen more sophisticated dissipation models we understand that there is a need for data-dependent filters. They should reduce instabilities, stemming from high-order approximations, and simultaneously enhancing the steepness of the shock front. In the following, we examine several nonlinear approaches in the context of image processing.

3.2 Nonlinear diffusion filters

As we have already seen linear diffusion filters are a very effective way to extract or reduce high frequency oscillations from a signal. However, due to their linearity the tendency to blur the signal is quite strong and leads to a smoothing of the gradients like edges, steps or corners which are intended to be enhanced or recovered. This leads to shape distortions, since smoothing over object boundaries can effect shape and localisation of the edge.

Therefore, there is a need to control the smoothing process which leads to a nonlinear and adaptive control of the diffusion filtering. This should be based on local properties of the signal in order to control the strength of the dissipation. The first formulation of such a nonlinear diffusion filter in image processing was given by Perona and Malik [88].

The basic idea is to modify the conductivity in the nonlinear diffusion equation

$$\partial_t u \quad = \quad \langle \nabla, c(u(t,x),t,x)\nabla u \rangle. \tag{3.5}$$

The conductivity is modified in such a way that it is low in regions where the gradient of u is high and vice versa.

In their original model, Perona and Malik proposed a diffusivity of the form

$$c(s) \quad = \quad \frac{1}{1+s^2/\lambda^2}, \qquad (\lambda > 0). \tag{3.6}$$

Here λ is a kind of contrast parameter which distinguishes regions requiring high or low diffusion.

In the following we discuss this model and the anisotropic extension based on an adapted diffusion tensor by Weickert [115, 116]. Later we present relations to techniques already developed in the area of numerical approximations of hyperbolic conservation laws.

The Perona-Malik model

The model equation

The diffusion model proposed by Perona and Malik integrates an adaptive control of the diffusion process in order to avoid the blurring of the signal by a linear dissipation model.

This is achieved by an inhomogeneous treatment of the filtering which reduces the strength of the diffusion at those locations which are considered to be edges (or shocks in the language of conservation laws). These are measured by the the steepness of the gradient $|\nabla u|^2$.

The Perona-Malik filter is modelled by the diffusion equation

$$\partial_t u \quad = \quad \mathrm{div}(c(|\nabla u|^2)\nabla u) \quad \text{in} \quad (0,T)\times\Omega \tag{3.7}$$

and initial data

$$u(0,x) \quad = \quad u_0(x).$$

Here, $c(s) : [0,+\infty) \to (0,+\infty)$ is the conductivity or diffusivity, which controls the dissipative behaviour of (3.7). In passing we note that choosing $c \equiv 1$ recovers the heat equation. If $c(s)$ is considered as a monotonically decreasing function with $c(0) = 1$ and $\lim_{s\to+\infty} c(s) = 0$ one notices that

- in regions where the magnitude of the gradient ∇u is small, the Perona-Malik model provides linear diffusion which results in isotropic diffusion like in the case of the heat equation.
- near discontinuities where the magnitude of the gradient of u is large, the regularisation is *stopped*, the conductivity cancels the diffusion and the discontinuity is preserved.

Thus, we see that the Perona-Malik model enhances edges because the diffusivity is adaptive or nonlinear.

Starting from a simple model in one space dimension and expanding the right hand side of (3.7) we derive

$$\partial_t u \quad = \quad c u_{xx} + c_x u_x.$$

This can be interpreted as a locally constant diffusion combined with a direction-dependent transport or drift. If a shock is on the left, $|\nabla u|$ increases, the diffusivity $c(|\nabla u|)$ decreases and the transport is proportional to $\partial_t u = -|\nabla u|$. The transport is into the shock which means backward diffusion. On the other side if the discontinuity $|\nabla u|$ decreases, c increases and the transport is like $\partial_t u = \nabla u$ into the shock. Around an edge backward diffusion occurs.

For the two-dimensional case – which for our purpose is the interesting one – we expand the divergence operator, i.e.

$$\begin{aligned}&\mathrm{div}(c(|\nabla u|^2)\nabla u)\\ &\quad = \quad 2(u_x^2 u_{xx} + u_y^2 u_{yy} + 2u_x u_y u_{xy})c'(|\nabla u|^2) + c(|\nabla u|^2)(u_{xx}+u_{yy}).\end{aligned}$$

If we define for each point where $|\nabla u| \neq 0$ the vectors

$$\underline{n} = \frac{\nabla u}{|\nabla u|} = \frac{1}{|\nabla u|}\begin{bmatrix} u_x \\ u_y \end{bmatrix}, \quad \underline{\xi} = \underline{n}^{\perp} = \frac{1}{|\nabla u|}\begin{bmatrix} -u_y \\ u_x \end{bmatrix},$$

the second derivatives of u in direction of $\underline{n}$ and $\underline{\xi}$ read as

$$\begin{aligned} u_{\underline{nn}} &= \underline{n}^T \nabla^2 u \underline{n} = \frac{1}{|\nabla u|^2}(u_x^2 u_{xx} + u_y^2 u_{yy} + 2u_x u_y u_{xy}), \\ u_{\underline{\xi\xi}} &= \underline{\xi}^T \nabla^2 u \underline{\xi} = \frac{1}{|\nabla u|^2}(u_x^2 u_{yy} + u_y^2 u_{xx} - 2u_x u_y u_{xy}). \end{aligned}$$

With $b(s) = c(s) + 2sc'(s)$ we are able to write (3.7) as

$$\partial_t u(t,x) = c(|\nabla u|^2) u_{\underline{\xi\xi}} + b(|\nabla u|^2) u_{\underline{nn}}. \tag{3.8}$$

Here one sees clearly that the nonlinear diffusion equation (3.7) may be interpreted as a weighted diffusion normal and tangential to the isolines of u, steered by the weighting coefficients c and b.

Well posedness

Naturally, since the task is to avoid smoothing of edges, one would prefer diffusion in direction tangential to the isolines rather than smoothing normal to them. Thus, it is reasonable to impose the condition

$$\lim_{s \to +\infty} \frac{b(s)}{c(s)} = 0,$$

or, equivalently,

$$\lim_{s \to +\infty} \frac{sc'(s)}{c(s)} = -\frac{1}{2}. \tag{3.9}$$

If the function $c(s)$ is assumed to be positive and with power growth, (3.9) implies $c(s) \approx 1/\sqrt{s}$ as $s \to +\infty$. Since we are concerned with a nonlinear diffusion equation, i.e variable coefficients, we have to assure that the problem (3.7) is still well posed, i.e. (3.7) remains parabolic.

Thus, rewriting the Perona-Malik equation (3.7) as

$$\partial_t u = a_{11}(|\nabla u|^2) u_{xx} + a_{12}(|\nabla u|^2) u_{xy} + a_{22}(|\nabla u|^2) u_{yy} \tag{3.10}$$

with coefficients

$$\begin{aligned} a_{11}(|\nabla u|^2) &= 2u_x^2 c'(|\nabla u|^2) + c(|\nabla u|^2), \\ a_{12}(|\nabla u|^2) &= 2u_x u_y c'(|\nabla u|^2), \\ a_{22}(|\nabla u|^2) &= 2u_y^2 c'(|\nabla u|^2) + c(|\nabla u|^2), \end{aligned}$$

(3.10) is parabolic if

$$\sum_{i,j=1,2} a_{ij}(|\nabla u|^2) \xi_i \xi_j \geq 0, \; \forall \xi \in \mathbb{R}^2. \tag{3.11}$$

This is equivalent to the condition that the matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$$

is positive definite, i.e. possesses positive eigenvalues. Hence the eigenvalues $z_{1,2}$ of $\mathbf{A}$ are given by

$$\mathsf{det}(z\mathbf{I} - \mathbf{A}) = \frac{a_{11} + a_{22}}{2} \pm (u_x^2 c' + u_y^2 c').$$

Thus, one obtains the eigenvalues

$$\begin{aligned} z_1 &= 2u_x^2 c' + 2u_y^2 c' + c, \\ &= 2sc' + c, \\ z_2 &= c, \end{aligned}$$

and with the demand $c > 0$ (3.11) is satisfied with

$$b(s) = z_1 = 2sc' + c > 0. \tag{3.12}$$

To summarise the assumptions we have to impose on the conductivity $c(s)$ the following conditions:

i) $c : [0, +\infty) \to (0, +\infty)$ decreasing,

ii) $c(0) = 1$, $c(s) \to 1/\sqrt{s}$ as $s \to +\infty$,

iii) $b(s) = c(s) + 2sc'(s) > 0$.

A function which satisfies these requirements is e.g.

$$c(s) = \frac{1}{\sqrt{1+s}}. \tag{3.13}$$

Edge enhancement

The idea behind the design of the Perona-Malik model (3.5) equipped with the diffusivity (3.13) was to choose the diffusion directions according to the variation of ∇u. This leads to a diffusion model which distinguishes between regions where smoothing takes place and regions where the solution remains unchanged.

For some purposes this approach is not sufficient, e.g. for edge enhancement where edges should be reconstructed and steepened. Thus, if we relax the restrictions (3.12) and impose a threshold parameter λ, such that $b(s) > 0$ for $s \leq \lambda$, and $b(s) < 0$ for $s > \lambda$, then (3.7) changes into a backward parabolic equation for $|\nabla u| > \lambda$, or equivalently into a smoothing-enhancing model.

We start from a simple one-dimensional model equation

$$\begin{aligned} \partial_t u &= \left[c(u_x^2(t,x))u_x(t,x)\right]_x \\ u(0,x) &= u_0(x) \end{aligned} \tag{3.14}$$

and examine the blurring/enhancing of an edge with respect to the diffusion coefficient $c(\cdot)$. Taking the derivative with respect to the space variable in (3.14) one derives formally with (3.12)

$$(\partial_t u)_x = \partial_t(u_x) = \partial_x([c(u_x^2)u_x]_x) = u_{xxx}b(u_x^2) + 2u_{xx}^2 b'(u_x^2). \tag{3.15}$$

If we assume a discontinuity at the location x at time t the second and third derivative of the solution u are nonpositive, i.e. $u_{xx}(t,x) = 0$ and $u_{xxx}(t,x) \leq 0$. So the sign of the change of the time derivative of u, i.e. the sign of (3.15) is determined by the sign of the coefficient b:

$$\operatorname{sign}(\partial_t u)_x = \operatorname{sign}(-b(u_x^2)(t,x)).$$

Thus, the change from forward to backward diffusion and so from smoothing to enhancement is controlled in the following way:

- For $b(u_x^2) > 0$ one has in (3.14) a forward diffusion process which corresponds to edge smoothing (blurring).
- For $b(u_x^2) < 0$ one has in (3.14) a backward diffusion process which corresponds to edge enhancing (sharpening).

Returning to the general two-dimensional diffusion model (3.7) we have already seen that one can rewrite it as in (3.8), which is the sum of weighted diffusion in direction normal and tangential to the isolines of u and so to possible discontinuities:

$$\partial_t u(t,x) = c(|\nabla u|^2)u_{\underline{tt}} + b(|\nabla u|^2)u_{\underline{nn}}.$$

Thus, if we intend to avoid smoothing and sharpen the discontinuity we need backward diffusion normal to the isolines, which imposes that the coefficient b is negative for large s, i.e.

$$b(s) = 2sc'(s) + c(s) < 0 \quad \text{for} \quad s \geq \lambda,$$

where λ is a given threshold or contrast parameter like in (3.6).

With similar arguments we are able to show that (3.14) equipped with (3.6) possesses the desired property and state the following

Lemma 3.3
The contrast parameter λ *separates the diffusion equation (3.14), rewritten as*

$$\partial_t u = \Phi'(u_x)u_{xx}$$

with the flux function *of the diffusivity (3.6), i.e.* $\Phi(s) := s\,c(s^2)$*, into a differential equation of*

- forward parabolic *type for* $|u_x| < \lambda$,
- backward parabolic *type for* $|u_x| > \lambda$.

It separates regions with low contrast – corresponding to forward diffusion – from regions with high contrast – according to backward diffusion.

Proof Assume a sufficiently smooth solution u and a point $x^\star$ where the gradient u_x^2 possess its maximum at time $t^\star$. As we have pointed out already this location is characterised by the conditions

$$u_x u_{xx} = 0, \qquad u_x u_{xxx} \leq 0.$$

Thus, on gets

$$\begin{aligned} \partial_t(u_x^2)(t^\star, x^\star) &= 2u_x \partial_x(u_t) \\ &= 2\Phi''(u_x) u_x u_{xx}^2 + 2\Phi'(u_x) u_x u_{xxx} \\ &= 2\Phi'(u_x) \underbrace{u_x u_{xxx}}_{\leq 0}. \end{aligned}$$

For the flux function $\Phi(s)$ of the diffusivity (3.6) we have

$$\begin{aligned} \Phi'(s) &= \tfrac{d}{ds}(s\, c(s)) \\ &= c(s) + s\, c'(s) \\ &= \frac{1}{1 + s^2/\lambda^2} + s \frac{-2s/\lambda^2}{(1 + s^2/\lambda^2)^2} \\ &= \frac{1 - s^2/\lambda^2}{(1 + s^2/\lambda^2)^2} \end{aligned}$$

We derive

$$\begin{aligned} \partial_t(u_x^2)(t^\star, x^\star) &\geq 0 \qquad |u_x(t^\star, x^\star)| > \lambda \\ &\qquad\qquad \text{for} \\ \partial_t(u_x^2)(t^\star, x^\star) &\leq 0 \qquad |u_x(t^\star, x^\star)| < \lambda \end{aligned}$$

with strict inequality for $u_x u_{xxx} \leq 0$. ∎

The contrast parameter λ plays the role of a threshold parameter. Since one has backward diffusion for $|\nabla u|$ small, the location of the edge will be kept. Furthermore the edges remain stable over much broader scales than with linear diffusion processes.

An open questions is still the existence of solutions for such kind of problems involving backward diffusion. In general, a backward diffusion process is an ill-posed problem. For diffusion models of the form (3.7) Kichenassamy [58] proved the following result:

Theorem 3.4
Assume that:

i) There exists a constant λ such that $b(s) > 0$ for $s < \lambda^2$ and $b(s) < 0$ for $s > \lambda^2$.

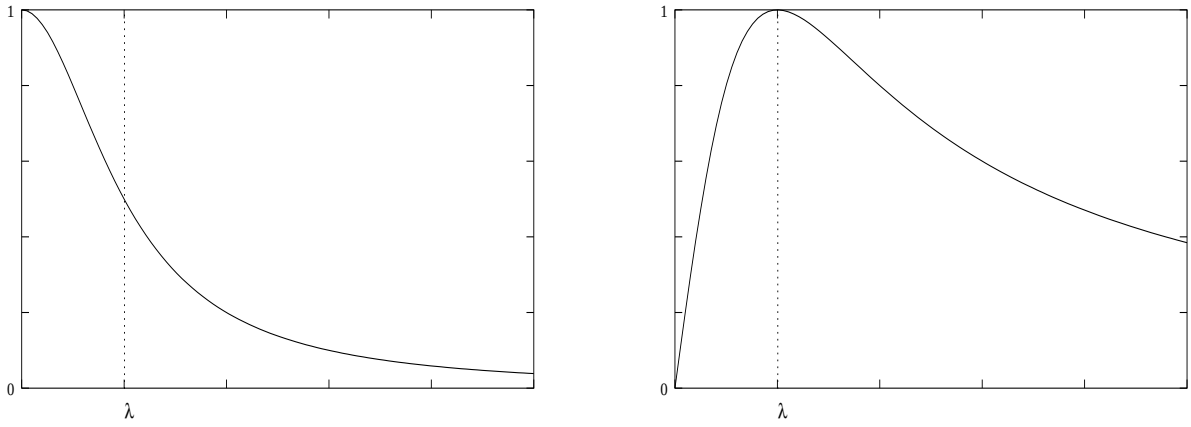

Figure 3.1: (a) Right Diffusivity $c(s) = \frac{1}{1+s^2/\lambda^2}$. (b) Left Flux function $\Phi(s) = \frac{s}{1+s^2/\lambda^2}$.

ii) Both functions, $c(s)$ and $b(s)$, tend to zero as $s \to +\infty$.

iii) (3.7) has a solution $u(t,x)$ satisfying $\lambda_1 \leq u_x(t,x) \leq \lambda_2 \quad \forall x \in [A,B], t \in [0,T]$ for some A, B and $\lambda_1 > \lambda$.

Then $u(t,x)$ is infinitely differentiable at $t = 0$, $\forall x \in (A,B)$. Therefore, if the initial image is not infinitely differentiable, there is no weak solution.

Remark 3.5
Hence, it is shown in [58], that the notion of a solution must be understood in the sense of measures. This means that the solution has to consist of continuous regions where the absolute value of the gradients is less than λ, separated by discontinuities which have infinite gradients, but measure zero.

Catté et al. [14] pointed out that for some diffusivities c inconsistencies arise with the scale-space theory developed in image processing. In order to obtain both existence and uniqueness, c must ensure that the flux $s\,c(s)$ is nondecreasing. If this requirement is not fulfilled, for some functions c with non-increasing flux a non-deterministic behaviour is observed. They propose the use of a diffusivity function based on the smoothed gradient, i.e. $c(|(\nabla G_\sigma * u)|)$ which avoids these difficulties.

Alvarez et al. [1] pointed out that under certain constraints, a natural choice for nonlinear diffusion is the equation

$$\begin{aligned} \partial_t u &= |\nabla u| \left\langle \nabla, \left(\frac{\nabla u}{|\nabla u|} \right) \right\rangle \\ &= |\nabla u| \kappa(u) \\ &= \partial^2_{\underline{\xi}\underline{\xi}} u, \end{aligned}$$

where $\kappa(u)$ denotes the curvature of level curves in u and $\underline{\xi}$ is the direction tangential to a level curve. This Ansatz leads to evolution of the level curves in the normal direction with velocity proportional to their curvature.

A similar approach was derived by Alvarez, Lions and Morel [2] which used a degenerated diffusion equation of the form

$$\partial_t u = g(|G_\sigma * \nabla u|)|\nabla u| \left\langle \nabla, \left(\frac{\nabla u}{|\nabla u|} \right) \right\rangle$$

Weickert pointed out, that although Perona and Malik classified their filter as anisotropic, it should by regarded as an isotropic one, since it utilises a scalar-valued diffusivity. He names a filter anisotropic, if the strength of diffusivity is given by a diffusion tensor.

TV-preserving models

We already have encountered the concept of the Total Variation in the previous chapter where it can be seen as a measure of simplicity of the solution since oscillations increase the total variation of a solution. This reason inspired Harten to introduce this concept to the design of numerical schemes in the area of fluid dynamics.

Osher and Rudin [93, 86] adapted these ideas into the area of image processing. They discuss the discrepancy between smoothing linear filters, which obey natural positivity requirements and non-smoothing linear filters which suffer from the problem of the so-called ringing phenomenon, of which the Gibbs oscillation is one example. To derive ringing free filters Osher and Rudin follow Harten and require positivity or monotonicity.

They proposed so-called shock filters which are nonlinear hyperbolic partial differential equations originating from ideas developed in the construction of numerical approximations for shock calculations. The idea is to sharpen a blurred edge in the data $u_0(x)$ by a backward diffusion controlled by the sign of the second derivative.

Shock filter

We start with some interesting investigations on the viscous Burgers' equation (1.18):

$$\partial_t u + u\partial_x u = \epsilon \partial_x^2 u. \tag{3.16}$$

The equation is balanced by convection and diffusion: the former tends to steepen the gradients while the latter smoothes the initial data. By the right control of the parameter ϵ this model could serve as a prototype filter. But some properties of (3.16) are not satisfying for the use as a filter:

- (3.16) is not symmetric, it will spread all the left facing profiles.
- The convective part let the solution propagate with different velocities depending on the different wave forms of the solution.
- For an image a two-dimensional filter is required.

Rudin [93] proposed an equation of the form

$$\partial_t u - F(|\nabla u|)\nabla u = \epsilon \Delta u,$$

to overcome these problems. He denoted filter algorithms of this kind as shock filter.

Here the symmetric form is achieved by taking the modulus of the gradient of u. The function F should control the propagation speed of the wave in such way, that it goes monotonically to zero since a derivative bound is exceeded. The new formulation overcomes the restriction to one dimension and since F depends on the gradient of u, rotational invariance is achieved.

In [86] Osher and Rudin improved their model for shock or TV-filters by calculating the restored image as the steady state solution of

$$\begin{aligned} \partial_t u &= -|\nabla u| F(\mathcal{L}(u)) \\ u(0, \underline{x}) &= u_0(\underline{x}), \end{aligned} \tag{3.17}$$

where F is a Lipschitz continuous function, obeying the following properties:

i) $F(0) = 0$,

ii) $\text{sign}(s)F(s) > 0, \quad s \neq 0.$

$\mathcal{L}$ is a second-order elliptic operator whose zero-crossing corresponds to edges, e.g. the Laplacian $\mathcal{L}(u) = \Delta u$ or the second-order directional derivative $\mathcal{L} = u_{\underline{\xi}\,\underline{\xi}}$ with $\underline{\xi}$ is again tangential to the level curves of u.

A typical and simple example in one space dimension is to take F as the identity, i.e. $F(u_{xx}) = u_{xx}$, so that (3.17) can be written as

$$\begin{aligned} u_t + (\text{sign}(u_x)u_{xx})u_x &= 0, \quad x \in \mathbb{R}_0^+ \times \mathbb{R} \\ u(0, x) &= u_0(x), \end{aligned}$$

This is a transport equation with nonlinear propagation speed $a(u) = \text{sign}(u_x)u_{xx}$. Since edges are regarded as maxima of $|u_x|$ where necessarily $u_{xx} = 0$, the propagation speed $a(u)$ serves as an edge detector for the diffusion model.

A simplified model of (3.17) is given by

$$\begin{aligned} u_t(t, x) &= -|u_x(t, x)|\text{sign}((u_0)_{xx}(x)), \\ u(0, x) &= u_0(x). \end{aligned} \tag{3.18}$$

One immediately sees that for areas where $u_x(t, x) > 0$ and $u_{xx}(t, x) > 0$, (3.18) acts like the transport equation $u_t + u_x = 0$ which is the desired motion for this points.

Aubert and Kornprobst [6] undertake a case study for (3.18) with initial data

$$u(0, x) = u_0(x) = \cos(x). \tag{3.19}$$

They show that with this initial data there exist a family of functions $u(t,x)_{t>0}$ such that for increasing t the limiting process tends to the step function $u(x) = (-1)^k$ for $(2k-1)\frac{\pi}{2} < x < (2k+1)\frac{\pi}{2}$. This means edge enhancing or edge formation at the inflection points of the initial data, i.e. $(u_0(x))_{xx} = 0$.

For a general one-dimensional model of (3.17) there is unfortunately no theoretical justification for this approach. It remains still to the class of ill-posed problems. Nevertheless, the numerical simulations presented by Osher and Rudin [86] are quite satisfying and based on these results they made the following

Conjecture 3.6
The evolution equation (3.17), with $u_0(x)$ continuous, has a unique solution that has jumps only at inflection points of $u_0(x)$ and for which the total variation in x of $u(t,x)$ is invariant in time, as well as the locations and values of local extrema.

The problem arising from the use of shock filters is the creation of shocks by fluctuation due to noise. The invention of new edges is clearly an unrequested feature and should be avoided. Alvarez and Mazorra [3] used a Gaussian-smoothed version of the elliptic operator $\mathcal{L}(u) = u_{\underline{\xi}\underline{\xi}}$ in (3.17), i.e. $\mathcal{L}(G_\sigma * u) = G_\sigma * u_{\underline{\xi}\underline{\xi}}$ and endowed this model with a mean curvature driven diffusion in order to remove noise. For the solution of their resulting semi-implicit finite-difference scheme they are able to prove uniqueness and a maximum-minimum principle.

From the viewpoint of fluid dynamics it is quite interesting that filter methods such as shock filters already occur in the context of shock calculations. Recently Bürgel and Sonar [11] showed that a simple form of a discrete filter algorithm for scalar conservation laws by Engquist, Lötstedt and Sjögreen [25] can be recast in the form of the general filter model (3.17) (see also [10]).

Anisotropic filter models

After the definition of anisotropic filters due to Weickert [116] all the filter algorithms presented above are isotropic since they are based on scalar-valued diffusivities rather than on a matrix valued form. Hence, the Perona-Malik model, often denoted as an anisotropic one, must be regarded after this definition as still isotropic. We follow this distinction and go to examine Weickert's approach [116] which is based on an adaptive diffusion tensor.

The basic filter

The structure of the filter algorithm is considered on a domain $\Omega \in \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$. The function $u(t, x_1, x_2)$ represents alternatively the grey values of an image or the value of a scalar function representing the conserved quantity of a conservation law. The continuous

anisotropic diffusion filter is given by the initial-boundary value problem

$$\begin{aligned} \partial_t u &= \operatorname{div}(\mathbf{D}\nabla u) \quad \text{for} \quad (t,\underline{x}) \in \mathbb{R}^+ \times \Omega, \\ u(0,\underline{x}) &= u_0(\underline{x}) \quad \text{for} \quad \underline{x} \in \Omega, \\ & \langle \mathbf{D}\nabla u, \underline{n} \rangle \quad \text{for} \quad (t,x_1,x_2) \in \mathbb{R}^+ \times \Gamma. \end{aligned} \tag{3.20}$$

Here, $\underline{n}$ is the outer normal. To reduce the sensitivity of the diffusion tensor $\mathbf{D} \in \mathbb{R}^{2\times 2}$ to noise, the tensor is based on the smoothed edge estimator ∇u_σ, i.e. on the convolved data

$$u_\sigma(t,\underline{x}) := (G_\sigma * u(t,\cdot))(\underline{x}), \qquad \sigma > 0.$$

Here, the diffusion tensor $\mathbf{D}$ should comprise the usual edge detector ∇u_σ, but also reveals more structural information from the data, e.g. preferred smoothing and enhancing orientations are extracted from the local data. To accomplish this Weickert used the so called structure tensor [29, 89], which is examined in the following.

The structure tensor

As one already has seen in the case of the Perona-Malik model there is a need for adaptive methods. They should take into account the local structure of the data like strong gradients etc. to overcome blurring effects like in linear filters.

To accomplish this goal, a structure tensor or interest operator is constructed. This operator should identify and analyse relevant features as well as measure the local coherence of significant structures in the given data. The basic tool for this construction is the vector-valued structure descriptor ∇u_σ derived from ∇u by smoothing with a Gaussian Kernel G_σ, where σ is within the length of typical small scale oscillation like white noise. The tensor $\mathbf{J}_0$ results from the tensor product of this structure descriptor

$$\begin{aligned} \mathbf{J}_0(\nabla u_\sigma) &:= \nabla u_\sigma \otimes \nabla u_\sigma \\ &:= \nabla u_\sigma^T \nabla u_\sigma \\ &= \begin{pmatrix} u_x^2 & u_x u_y \\ u_y u_x & u_y^2 \end{pmatrix}. \end{aligned}$$

This matrix has an orthonormal basis of eigenvectors $\underline{v}_1$, $\underline{v}_2$ with

$$\begin{aligned} \underline{v}_1 &\parallel \nabla u_\sigma, \\ \underline{v}_2 &\perp \nabla u_\sigma. \end{aligned}$$

The corresponding eigenvalues are $|\nabla u_\sigma|^2$ and 0.

As before, this structure descriptor is quite sensitive to noise like spurious oscillations which occurs naturally in the vicinity of discontinuities. Again, to limit the influence of noise, the orientation information from the original structure tensor is averaged by componentwise convolving with a Gaussian kernel G_ρ, i.e.

$$\mathbf{J}_\rho(\nabla u_\sigma) := G_\rho * (\nabla u_\sigma \otimes \nabla u_\sigma), \qquad \rho \geq 0,$$

is computed. The width ρ again is a measure of the averaging region. In practice, we solve the discrete heat equation componentwise for the structure tensor.

A simple computation with the matrix

$$\mathbf{J}_\rho(\nabla u_\sigma) \quad = \quad \begin{bmatrix} j_{11} & j_{12} \\ j_{21} & j_{22} \end{bmatrix}$$

reveals the eigenvalues

$$\lambda_{1,2;\rho} = \frac{1}{2}\left(j_{11} + j_{22} \pm \sqrt{(j_{11} - j_{22})^2 + 4j_{12}^2}\right). \tag{3.21}$$

They correspond to the eigenvectors

$$\begin{aligned} \underline{v}_{1;\rho} &= \begin{bmatrix} 2j_{12} \\ j_{22} - j_{11} + \sqrt{(j_{11} - j_{22})^2 + 4j_{12}^2} \end{bmatrix}, \\ \underline{v}_{2;\rho} &= \begin{bmatrix} j_{11} - j_{22} - \sqrt{(j_{11} - j_{22})^2 + 4j_{12}^2} \\ 2j_{12} \end{bmatrix}, \end{aligned} \tag{3.22}$$

which again are orthogonal. The parameter σ in the pre-smoothing processing is called the local scale or noise scale, because the process of pre-smoothing neglects all scales smaller than $\mathcal{O}(\sigma)$. In contrast, the parameter ρ is the integration scale indicating the size of the subregions in which the orientation of the numerical solution is analysed. The eigenvalues $\lambda_{1,2;\rho}$ moreover serve as descriptors of local structure. Constant solutions are characterised by $\lambda_{1;\rho} = \lambda_{2;\rho} = 0$, while the quantity

$$(\lambda_{1;\rho} - \lambda_{2;\rho})^2 = (j_{11} - j_{22})^2 + 4j_{12}^2$$

becomes large for anisotropic structures. In the language of image processing $(\lambda_{1;\rho} - \lambda_{2;\rho})^2$ is considered as a measure of local coherence.

With some assumptions on the diffusion tensor $\mathbf{D}(\mathbf{J}_\rho(\nabla u_\sigma)) = (d_{ij})$, Weickert is able to prove the following important result, concerning well-posedness, regularity and an extremum principle [116]:

Theorem 3.7
Consider the initial value problem (3.20), where the diffusion tensor $\mathbf{D}(\mathbf{J}_\rho(\nabla u_\sigma))$ satisfies the following properties:

i) Smoothness and symmetry:
$D \in C^\infty(S^2, S^2)$, where S^2 denotes the set of symmetric matrices.

ii) Uniform positive definiteness:
For all $w \in L^\infty(\Omega, \mathbb{R}^2)$ with $|w(x)| \leq K$ on $\overline{\Omega}$, there exists a positive lower bound $\nu(K)$ for the eigenvalues of $\mathbf{D}(\mathbf{J}_\rho(w))$.

Then this problem has an unique solution $u(t, \underline{x})$ in the distributional sense which satisfies

$$\begin{aligned} u \in C([0,T]; L^2(\Omega)) \cap L^2(0,T; H^1(\Omega)), \\ \partial_t u \in L^2(0,T; H^1(\Omega)). \end{aligned}$$

Furthermore, one has $u \in C^\infty((0,T] \times \overline{\Omega})$. *The solution depends continuously on the initial data* $u_0(\underline{x})$ *with respect to* $\|\cdot\|_{L^2(\Omega)}$, *and it satisfies the extremum principle:*

$$\inf_\Omega u_0(\underline{x}) \leq u(t,\underline{x}) \leq \sup_\Omega u_0(\underline{x}).$$

Diffusion tensor

Since the diffusion tensor should render the local structure of the data and steer the diffusion by a chosen diffusivity, it is clear that dissipation tensor $\mathbf{D}$ makes use of the information contained in the structure tensor. Weickert chooses the diffusion tensor to be

$$\mathbf{D}(\nabla u_\sigma) := \mathbf{V}_\rho \mathbf{L} \mathbf{V}_\rho^{-1} \tag{3.23}$$

where $\mathbf{V}_\rho$ contains the eigenvectors of $\mathbf{J}_\rho$ and $\mathbf{L} = \operatorname{diag}(l_1, l_2)$ is a diagonal matrix whose entries we have to choose properly. In order to recover shocks (or, equivalently, in order to enhance edges), the diffusivity l_1 perpendicular to edges should be reduced if the contrast $\lambda_{1;\rho}$ is high. This can be achieved by an anisotropic regularisation of the Perona-Malik model (3.7):

$$\begin{aligned} l_1 &= \vartheta(\lambda_{1;\rho}) \\ l_2 &= 1 \\ \vartheta(s) &= \begin{cases} 1 & s \leq 0, \\ \text{for} & \\ 1 - \exp\left(\frac{-C_m}{(s/\lambda)^m}\right) & s > 0. \end{cases} \end{aligned} \tag{3.24}$$

The values of m and C_m are chosen in such a way, that the flux function $\Phi(s) := s\vartheta(s)$ is increasing in an interval $s \in [0,\lambda]$ and decreasing in $s \in]\lambda, \infty[$. These choices depend on a one-dimensional analysis of the classical Perona-Malik model and we refer the reader to Weickert's book for details. A good choice is $m = 4$ and thus $C_4 = 3.31488$. The parameter λ can then be chosen freely.

Numerical analysis is the study of algorithms for the problems of continuous mathematics.

N. L. Trefethen
("The definition of numerical analysis" [108])

4 Discrete filters for scalar conservation laws

We already have seen in the second chapter that a numerical algorithm approximating a conservation law needs a dose of built-in artificial dissipation to stabilise the scheme. The question is how to choose and steer these diffusion terms. Necessarily they are nonlinear since the class of linear diffusion filters which we have discussed in the foregoing chapter is monotone but highly dissipative. This was shown by the analysis of the most prominent representative of this class, the Lax-Friedrichs scheme (2.27).

Since we already have discussed several approaches for nonlinear anisotropic diffusion stemming from image processing in the foregoing it seems quite reasonable to check whether these algorithms can be adapted to the needs arising from applications in the context of Computational Fluid Dynamics. This paragraph is concerned with this approach.

4.1 The basic filter

In the development of higher order numerical methods for conservation laws there are in principle two ways for the design of such schemes. The first path traces back to ideas by von Neumann[1] and Richtmyer [111] and were revived by Jameson, Schmidt and Turkel [56] at the beginning of the 1980's, starting directly from the construction of the numerical viscosity coefficients. This construction is very difficult due to the inherent nonlinearity of the problem and a hard obtainable well-behaving dissipation. Moreover, this method relies on some special switches and user-defined parameters which demands a deep knowledge of the problem. However, the code of Jameson et al. still belongs to the successful tools of an engineer and is known for its flexibility as well as its stability.

The other track started in the mid 80's and is related to the development of TVD and ENO schemes. Here one starts from a low order monotone method, recovers the solution (with MUSCL or ENO techniques [44, 48, 49, 109, 110]) and inserts the recovered solution into the low order flux function. The recipe is quite general and the most successful schemes, like the mentioned TVD and ENO approaches, rely on this construction. Although the process of recovery itself is not such easy, one knows quite well how this can be done even on unstructured grids. In this class of methods one does not even see the numerical dissipation

[1] John von Neumann, Berlin, Göttingen, Princeton (1903 – 1957)

of the method, which may seem a big advantage over the Jameson-type method.

However, we try to proceed along the line of Jameson and his co-workers and construct a reasonable dissipation filter directly based on the algorithms from image processing, namely the anisotropic diffusion algorithm developed by Weickert [116].

This means we construct a diffusion matrix

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \tag{4.1}$$

similar to the diffusion tensor following Weickert's recipe. The additional feature for the filter design is given by the fact that in general the non-diagonal coefficients $d_{12}, d_{21} \neq 0$, i.e. cross diffusion takes places. Thus, the resulting method has to be regarded as a genuinely multidimensional one in contrast to classical one-dimensional approaches like the Strang splitting [103].

For the design of the dissipation filter we require some properties which the model has to satisfy in order to turn the underlying Lax-Wendroff method into a high resolution method:

- Stabilise the underlying second-order scheme.
- Remove the small scale oscillations occurring in the vicinity of discontinuities due to the instability of second-order methods.
- Fulfil this task by enhancing the shock, i.e. reducing the diffusion perpendicular to the discontinuity and adding the major amount parallel to it in order to keep the gradient steep.
- Keep the second-order accuracy in the vicinity of the shock.

Since classical high resolution methods like schemes including flux or slope limiter reduce at shocks to first order accuracy it seems reasonable for a new approach to apply as the first step the second-order oscillatory Lax-Wendroff scheme which controls the convective part of the numerical scheme. This ensures second order accuracy in space and time. The difficult task that remains is to remove the oscillations by keeping the accuracy.

In the second step, we may view the resulting numerical solution as a 'picture' with possible noisy edges coming from the oscillations of the Lax-Wendroff scheme at shocks (stemming from the Gibbs phenomenon). We apply the described algorithms from image processing tailored to the requirements for Computational Fluid Dynamics, and enhance the shock by removing the wiggles along the discontinuity and conserve the accuracy in this area.

We start with a model equation

$$\partial_t u + \partial_{x_1} f(u) + \partial_{x_2} g(u) = 0 \tag{4.2}$$

on $\Omega := [0,1]^2$. Cauchy data $u_0(\underline{x}) = u(0, \underline{x})$ as well as boundary data respecting the characteristic directions are given.

The convective step

In two dimension, there are a variety of different implementations of the Lax-Wendroff scheme like the two-step variant proposed by Richtmyer and Morton [90] or the modification by Eilon, Gottlieb and Zwas [23]. We use the canonical extension of the original algorithm by Lax and Wendroff [70] which can be found in the book of Shokin [97]:

$$\begin{aligned}
&\frac{U^{i,j}-U_{i,j}}{\Delta t}+\frac{F_{i+1,j}-F_{i-1,j}}{2h_1}+\frac{G_{i,j+1}-G_{i,j-1}}{2h_1}\\
&= \frac{\lambda_1}{2}\left[A_{i+1/2,j}\left(\frac{F_{i+1,j}-F_{i,j}}{h_1}+\frac{G_{i+1/2,j+1/2}-G_{i+1/2,j-1/2}}{h_2}\right)\right.\\
&\qquad \left.-A_{i-1/2,j}\left(\frac{F_{i,j}-F_{i-1,j}}{h_1}+\frac{G_{i-1/2,j+1/2}-G_{i-1/2,j-1/2}}{h_2}\right)\right]\\
&\quad +\frac{\lambda_2}{2}\left[B_{i,j+1/2}\left(\frac{F_{i+1/2,j+1/2}-F_{i-1/2,j+1/2}}{h_1}+\frac{G_{i,j+1}-G_{i,j}}{h_2}\right)\right.\\
&\qquad \left.-B_{i-1/2,j}\left(\frac{F_{i+1/2,j-1/2}-F_{i-1/2,j-1/2}}{h_1}+\frac{G_{i,j}-G_{i,j-1}}{h_2}\right)\right],
\end{aligned} \tag{4.3}$$

with $\lambda_1 = \Delta t/\Delta x_1 = \Delta t/h_1$, $\lambda_2 = \Delta t/\Delta x_2 = \Delta t/h_2$ and

$$A_{i\pm1/2,j} := f'\left(\frac{U_{i\pm1,j}+U_{i,j}}{2}\right) \quad \text{and} \quad B_{i,j\pm1/2} := g'\left(\frac{U_{i,j\pm1}+U_{i,j}}{2}\right).$$

The amplification factor of the linearised scheme is given by

$$\begin{aligned}
\mathsf{A} &= 1 - i(\lambda_1 A \sin\Theta_1 + \lambda_2 B\cos\Theta_2)\\
&\quad -\left[\lambda_1^2 A^2(1-\cos\Theta_1)+\lambda_2^2 B^2(1-\cos\Theta_2)+\lambda_1\lambda_2 AB\sin\Theta_1\Theta_2\right]
\end{aligned}$$

which leads to the CFL-condition

$$\frac{\Delta t}{h}\max_{u\in\Omega}[A(u),B(u))] \le \frac{1}{2\sqrt{2}} \tag{4.4}$$

for $h_1 = h_2 = h$ [13].

Although the scheme is stable obeying the condition (4.4), it produces unphysical oscillations which can be seen in Figure 4.1.

The dissipation step

Since we are interested in a dissipation algorithm which makes use of a direction-dependent diffusion parallel to shocks, we are going to construct the dissipation filter after the recipe given by Weickert [116] for anisotropic diffusion in image processing. Thus, the second step of the algorithm will be a discrete form of the dissipation model

$$\partial_t u = \operatorname{div}[\mathbf{D}(u)\nabla u] \tag{4.5}$$

with dissipation matrix

$$\mathbf{D} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}. \tag{4.6}$$

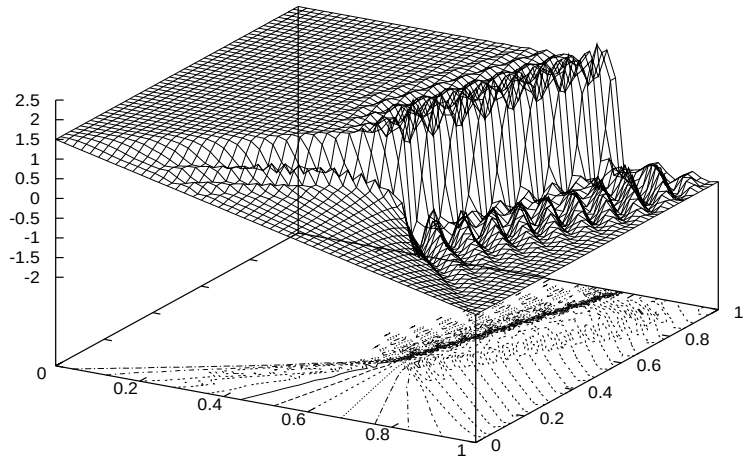

Figure 4.1: Numerical solution of the Lax-Wendroff scheme (4.3) without filtering

The dissipation matrix should reveal an isotropic diffusion filter for smooth regions while in the vicinity of a shock the filter should become anisotropic, i.e. prefer dissipation parallel to the discontinuity in order to avoid blurring.

Thus, we need a discrete data analysis to detect the local flow structure which is gained from information inside the structure tensor, i.e. its eigenvectors and eigenvalues. Furthermore, we need information about the diffusion strength and have to consider the need of the integration scales in order to average the data used to build the structure tensor and the diffusion matrix.

The structure tensor

There are already different approaches to detect the direction of the local flow structures based on a discrete data analysis. For instance, for the Euler equations the method of Murman and Cole [84] solves the transonic potential equation in supersonic regions by replacing derivatives in streamwise direction by upwind difference approximations and in the normal direction to the streamlines by central difference approximations. This is easy as long as the computational grid is already approximately aligned with the streamline and normal direction, but difficult otherwise.

Jameson [55] and Davis [22] overcome this problem by introducing a local coordinate system based on the streamline and normal directions. Davis explicitly designed a method to resolve shocks by using different numerical flux functions in directions normal and tangential to shocks.

Remark 4.1
Since Davis has developed a method for the Euler equations he uses the flow velocities as indicators for a discontinuity because a shock is normal to the jump in the velocity (see Davis

[22] and e.g. Liepmann and Roshko [74]). With the one-sided forward difference operators $D^+_{x_1}U_{i,j} := (U_{i+1,j} - U_{i,j})$, $D^+_{x_2}U_{i,j} := (U_{i,j+1} - U_{i,j})$ *the angle* Θ^x *is used to compute rotated numerical fluxes at the cell interface passing through the point* $(x_{i+1/2}, y_i)$ *is computed by*

$$\begin{aligned}\Theta^x_{i+1/2,j} &:= \arctan\left(\frac{D^+_{x_1}U_{i,j}}{D^+_{x_2}V_{i,j}}\right)\\ &= \arctan\left(\frac{U_{i+1,j} - U_{i,j}}{V_{i,j+1} - V_{i,j}}\right)\end{aligned}$$

for the velocity vector $\underline{v} = (u, v)^T$.

Similarly, the angle Θ^y *is used to compute rotated numerical fluxes at the cell interface which passes through* $(x_i, y_{j+1/2})$ *is given by*

$$\begin{aligned}\Theta^y_{i,j+1/2} &:= \arctan\left(\frac{D^+_{x_2}U_{i,j}}{D^+_{x_1}v_{i,j}}\right)\\ &= \arctan\left(\frac{U_{i,j+1} - U_{i,j}}{V_{i+1,j} - V_{i,j}}\right).\end{aligned}$$

The angle $\Theta^{x/y}$ *may vary widely in smooth parts of the flow field. To overcome this problem, Davis proposed to use weighted averages of the differences similar to Weickert's method. We will discuss this fact in the following.*

From the – possibly pre-smoothed – data u_σ, the structure tensor is computed from

$$\mathbf{J}_0(\nabla U_{i,j;\sigma}) := \nabla U_{i,j;\sigma}\nabla U^T_{i,j;\sigma}$$

which is symmetric positive semidefinite. The discrete nabla operator is computed with centred difference operators $D^0_{x_1}U_{i,j} := (U_{i+1,j} - U_{i-1,j})$, $D^0_{x_2}U_{i,j} := (U_{i,j+1} - U_{i,j-1})$ so that it reads for the cell $\mathcal{C}_{i,j}$ as

$$\begin{aligned}\nabla U_{i,j;\sigma} &:= \begin{bmatrix}\dfrac{D^0_{x_1}U_{ij;\sigma}}{2h_1}\\ \dfrac{D^0_{x_2}U_{ij;\sigma}}{2h_2}\end{bmatrix}\\ &= \begin{bmatrix}\dfrac{U_{i+1,j;\sigma} - U_{i-1,j;\sigma}}{2h_1}\\ \dfrac{U_{i,j+1;\sigma} - U_{i,j-1;\sigma}}{2h_2}\end{bmatrix}.\end{aligned}$$

An easy calculation reveals the eigenvalues of $\mathbf{J}_0$

$$\lambda_1 = |\nabla U_{i,j;\sigma}|^2, \quad \lambda_2 = 0,$$

corresponding to the eigenvectors

$$\underline{v}_1 = \nabla U_{i,j;\sigma}, \quad \underline{v}_2 = \nabla^\perp U_{i,j;\sigma},$$

where

$$\nabla^\perp U_{i,j;\sigma} := \begin{bmatrix}-\dfrac{U_{i,j+1;\sigma} - U_{i,j-1;\sigma}}{2h_2}\\ \dfrac{U_{i+1,j;\sigma} - U_{i-1,j;\sigma}}{2h_1}\end{bmatrix}.$$

In fact, the eigenvectors of the structure tensor define the direction parallel to, and across an edge, respectively. Thus, the structure tensor plays the role of an operator designed for structure detection, which mirrors the information about orientation and magnitude of high and low contrast in the eigenvectors and eigenvalues.

Since the structure tensor is symmetric positive semidefinite one has the splitting

$$\mathbf{J}_0(\nabla U_{i,j;\sigma}) \;=\; \mathbf{V}\,\mathbf{\Lambda}\,\mathbf{V}^{-1},$$

where $\mathbf{V} = (\underline{v}_1, \underline{v}_2)$ is the transformation matrix containing the eigenvectors and $\mathbf{\Lambda}$ is nothing but $\mathrm{diag}(\lambda_1, \lambda_2)$. Note, that due to the symmetry of $\mathbf{J}_0$ we have $\mathbf{V}^T = \mathbf{V}^{-1}$.

Computing the angle

$$\Theta \;:=\; \arctan\left(\frac{D^0_{x_2} U_{i,j;\sigma}}{D^0_{x_1} U_{i,j;\sigma}}\right) \;=\; \arctan\left(\frac{U_{i,j+1;\sigma} - U_{i,j-1;\sigma}}{U_{i+1,j;\sigma} - U_{i-1;j}}\right)$$

we can rewrite the structure tensor as a function of Θ and – since $\|\underline{v}_1\| = \|\underline{v}_2\| =: \|\underline{v}\|$ – of $\|\underline{v}\|$:

$$\begin{aligned}\mathbf{J}_0(\Theta, \|\underline{v}\|) \;&=\; \begin{pmatrix} \|v\|\cos\Theta & -\|v\|\sin\Theta \\ \|v\|\sin\Theta & \|v\|\cos\Theta \end{pmatrix} \begin{pmatrix} \|v\|^2 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \|v\|\cos\Theta & \|v\|\sin\Theta \\ -\|v\|\sin\Theta & \|v\|\cos\Theta \end{pmatrix} \\ &=\; \|\underline{v}\|^4 \begin{pmatrix} \cos^2\Theta & \cos\Theta\sin\Theta \\ \sin\Theta\cos\Theta & \sin^2\Theta \end{pmatrix}.\end{aligned}$$

Remark 4.2
Here one clearly sees the relation to the pioneering work of Davis [22]. Similar to his method a transformation of the basis vectors takes place which is equivalent to the rotation with the angle Θ. In contrast to Davis' approach the rotation is computed for the whole cell and is only applied to the diffusion model which will be weighted by nonlinear diffusivities.

Averaging the structure information If one is interested in averaging the orientation information in order to reduce influences of strongly varying gradients, componentwise convolution of $\mathbf{J}_0(\nabla u_\sigma)$ with a Gaussian kernel of width ρ, G_ρ is appropriate. Applying this method reveals the averaged structure tensor

$$\mathbf{J}_\rho(\nabla u_\sigma) \;:=\; G_\rho * (\nabla u_\sigma \nabla u_\sigma^T),$$

where the integration scale ρ is a measure of the averaged region over which the orientation is analysed. In practice, on the discrete level, the convolution is again accomplished by solving the finite difference formula for the heat equation for each component.

The matrix

$$\mathbf{J}_\rho(\nabla U_{i,j;\sigma}) \;=:\; \begin{bmatrix} j_{11} & j_{12} \\ j_{12} & j_{22} \end{bmatrix} \tag{4.7}$$

possesses eigenvalues $\lambda_{1,2;\rho}$ and eigenvectors $\underline{v}_{1;\rho}, \underline{v}_{2;\rho}$ introduced in (3.21) and (3.22), respectively.

The eigenvalues $\lambda_{1,2;\rho}$ again can be interpreted as descriptors of local structure. Constant solutions are characterised by $\lambda_{1;\rho} = \lambda_{2;\rho} = 0$, while the quantity

$$(\lambda_{1;\rho} - \lambda_{2;\rho})^2 = (j_{11} - j_{22})^2 + 4j_{12}^2,$$

introduced as the measure of local coherence, becomes large for anisotropic structures.

Averaging scale σ The averaging or smoothing parameter σ and ρ are introduced into the concept of feature detection by the construction of the structure tensor to analyse the data over the subregions which size is given by this scales. The local scale or noise scale σ which is used in the pre-smoothing process, is used to make the computation of the discrete operator $\nabla U_{i,j;\sigma}$ insensitive to oscillations and irrelevant details smaller than $\mathcal{O}(\sigma)$. Thus, in the area of approximative solutions of conservation laws one would expect the local scale as $\sigma = \mathcal{O}(2\max[h_1, h_2])$.

But necessarily, the smoothing process not only leads to reduction of the oscillation but to smoothing of the general structure. Due to the nonlinearity of conservation laws in general, one is faced with a strong dependence of the solution to changes in the data. Thus, it is the question whether averaging inside the feature detector is desirable at all, since some information will be lost.

One way to solve this problem is to use the 'best' first order method, i.e. the Godunov scheme. If we consider to start from a monotone solution U for the previous time level t^n we use the Lax-Wendroff scheme represented by the operator $\mathcal{C}^{LW}$ for the convective step and use the Godunov operator $\mathcal{S}^{God}$ applied to the original data at time level t^n for the structure tensor. Since the Godunov scheme is a monotone method, i.e. does not create new extrema with respect to monotone data, one gets oscillation-free data for the construction of the structure tensor, i.e.

$$U_{i,j;\sigma} \quad := \quad \mathcal{S}^{God} U_{i,j}. \tag{4.8}$$

Another possible approach is to use the second-order data in smooth regions and the first-order data, computed with the Godunov method, at cells where new extrema are created by the Lax-Wendroff formula:

$$U_{i,j;\sigma} \quad = \quad \begin{cases} \mathcal{S}^{God} U_{i,j} & \text{for} \quad \left(\dfrac{D_{x_1}^+ \mathcal{C}^{LW} U_{i,j}}{D_{x_1}^- \mathcal{C}^{LW} U_{i,j}}\right) \cdot \left(\dfrac{D_{x_1}^+ U_{i,j}}{D_{x_1}^- U_{i,j}}\right) < 0 \\ & \text{or} \quad \left(\dfrac{D_{x_2}^+ \mathcal{C}^{LW} U_{i,j}}{D_{x_2}^- \mathcal{C}^{LW} U_{i,j}}\right) \cdot \left(\dfrac{D_{x_2}^+ U_{i,j}}{D_{x_2}^- U_{i,j}}\right) < 0 \\ \mathcal{C}^{LW} U_{i,j} & \text{else} \end{cases} .$$

Here, $D_{x_1}^- U_{i,j} := (U_{i,j} - U_{i-1,j})$, $D_{x_2}^- U_{i,j} := (U_{i,j} - U_{i,j-1})$ are the one-sided backward difference operators. The products of the quotients of the consecutive gradients

$$\left(\frac{D_{x_1}^+ \mathcal{C}^{LW} U_{i,j}}{D_{x_1}^- \mathcal{C}^{LW} U_{i,j}}\right) \cdot \left(\frac{D_{x_1}^+ U_{i,j}}{D_{x_1}^- U_{i,j}}\right) = \left(\frac{U_{i+1,j}^{LW} - U_{i,j}^{LW}}{U_{i,j}^{LW} - U_{i-1,j}^{LW}}\right) \cdot \left(\frac{U_{i+1,j} - U_{i,j}}{U_{i,j} - U_{i-1,j}}\right)$$

and

$$\left(\frac{D_{x_2}^{+} \mathcal{C}^{LW} U_{i,j}}{D_{x_2}^{-} \mathcal{C}^{LW} U_{i,j}} \right) \cdot \left(\frac{D_{x_2}^{+} U_{i,j}}{D_{x_2}^{-} U_{i,j}} \right) = \left(\frac{U_{i,j+1}^{LW} - U_{i,j}^{LW}}{U_{i,j}^{LW} - U_{i,j+1}^{LW}} \right) \cdot \left(\frac{U_{i,j+1} - U_{i,j}}{U_{i,j} - U_{i,j-1}} \right)$$

serve as sensors for the creation of new extrema inside the cell $\mathcal{C}_{i,j}$. If the ratio of the consecutive gradients

$$\left(\frac{U_{i+1,j} - U_{i,j}}{U_{i,j} - U_{i-1,j}} \right)$$

is negative, we assume an extremum in cell $\mathcal{C}_{i,j}$. If the consecutive gradients of the Lax-Wendroff solution U^{LW} and of the old time level U have different signs then clearly a new extremum in this cell was created.

Averaging scale ρ The same questions exists for the integration parameter ρ, namely whether the averaging of the structure tensor is an adequate method or not. In our numerical examples we did not discover a remarkable difference in the solutions. Nevertheless there may be situations where the integration scale plays a remarkable role, e.g. if strong changes in the direction information – corner like structures – take place. Thus, even if the process of averaging is not used in our algorithms – also because smoothing is a quite expensive process with respect to computation time – it is inserted as an element in the list describing the construction of the structure tensor, in order to derive a complete design concept for the filter.

Remark 4.3
As already mentioned in the context of the computation of the direction information, the approach of Davis also uses averaged data to compute the angle Θ. He used averaged differences of the form

$$\delta_{x_1} U_{i,j} := \left(\frac{\Delta_{x_1} U_{i+1,j} + 4\Delta_{x_1} U_{i,j} + \Delta_{x_1} U_{i-1,j}}{6} \right)$$

where

$$\Delta_{x_1} U_{i,j} := \left(\frac{D_{x_1}^{-} U_{i,j+1} + 4 D_{x_1}^{-} U_{i,j} + D_{x_1}^{-} U_{i,j-1}}{6} \right).$$

One sees that for the computation of the gradient averages in the direction of the gradient are chosen biased by distance weights. These averaged values itself are computed by the same approach using averaged information perpendicular to the direction of the gradient. This ought to make the direction information insensitive to noise and small scale oscillations.

The diffusion matrix D

As the last step of the algorithm, after the relevant features of the data are detected by the structure tensor, we have to use this information in order to derive a dissipation model. This is nothing more than a sophisticated version of the anisotropic diffusion filter given

by Weickert. 'Sophisticated' is meant here in the way that we have to adapt the model to approximative solution for conservation laws and hence weight the diffusion strength in a suitable manner.

We normalise the set of eigenvectors (3.22) such that the matrix $\mathbf{V}_\rho = [\underline{v}_{1;\rho}, \underline{v}_{2;\rho}]$ contains a set of orthonormal eigenvectors instead of only orthogonal ones. This is necessary since the matrices $\mathbf{V}_\rho$ and $\mathbf{V}_\rho^T$ should only rotate the coordinate system rather than stretch.

Thus, the Ansatz for the diffusion matrix is

$$\mathbf{D} \quad := \quad \mathbf{V}_\rho \mathbf{L} \mathbf{V}_\rho^T, \tag{4.9}$$

with the diagonal matrix $\mathbf{L} = \mathrm{diag}(l_1, l_2)$ where the diffusion coefficients l_1, l_2 have to be determined properly. If one introduces the angle

$$\Theta_\rho \quad := \quad \arctan\left(\frac{v_{1_\rho}^2}{v_{1_\rho}^1}\right)$$

the splitting (4.9) can be written in the form

$$\begin{aligned}
\mathbf{D} &= \begin{bmatrix} \cos\Theta_\rho & -\sin\Theta_\rho \\ \sin\Theta_\rho & \cos\Theta_\rho \end{bmatrix} \begin{bmatrix} l_1 & 0 \\ 0 & l_2 \end{bmatrix} \begin{bmatrix} \cos\Theta_\rho & \sin\Theta_\rho \\ -\sin\Theta_\rho & \cos\Theta_\rho \end{bmatrix} \\
&= \begin{bmatrix} l_1\cos^2\Theta_\rho - l_2\sin^2\Theta_\rho & (l_1 - l_2)\sin\Theta_\rho\cos\Theta_\rho \\ (l_1 - l_2)\sin\Theta_\rho\cos\Theta_\rho & l_1\sin^2\Theta_\rho + l_2\cos^2\Theta_\rho \end{bmatrix} \\
&= \begin{bmatrix} a & b \\ b & c \end{bmatrix}.
\end{aligned}$$

This representation makes quite clear that we have constructed a dissipation matrix which rotates any vector with the angle Θ_ρ into the new coordinate system ξ, η. Here, the new coordinate axes are parallel and perpendicular to the shock, respectively. The components of the transformed vector are weighted by the diffusion coefficients and transformed back to the original coordinate system x_1, x_2. Thus, in order to restrict diffusion across a relevant feature like a discontinuity we have to choose the diffusion coefficients in a suitable manner.

The diffusion model If the transformation into the new coordinate system is accomplished one has to choose a suitable dissipation model which takes the local data into account. Since the goal is to enhance the shock and conserve the high order of the underlying scheme it seems reasonable to control the diffusion applied to the discontinuity. Thus, we reduce the diffusion in the vicinity of the shock to an amount which is strong enough to avoid oscillations and which is small enough to preserve the order of the scheme. The diffusion applied parallel to the shock should support this aim and can be applied in full strength.

To recover the shock we choose the diffusivities l_1, l_2 of the diagonal matrix $\mathbf{L}$ in accordance with (3.24).

Discretising the diffusion equation

After deriving the nonlinear anisotropic diffusion equation which will sharpen the shocks this equation needs now to be discretised. Since we do not have a theory of a truly discrete

diffusion equation to start with but a partial differential equation the discretisation process may result in instabilities if done in a naive way.

As was shown by Weickert [116] there is always a finite difference stencil such that the resulting discretisation leads to a stable scheme. Moreover, Weickert was able to prove that three directions suffice to discretise the anisotropic diffusion and the proof is constructive. We do not want to go into the details of Weickert's work but give a suitable discretisation of $\operatorname{div}(\mathbf{D}(u)\nabla u)$.

Following Weickert's recipe, one gets

$$\operatorname{div}(\mathbf{D}(U_{i,j;\delta})\nabla U_{i,j;\delta}) = \sum_{k=-1}^{1}\sum_{\ell=-1}^{1} C_{i+k,j+\ell}U_{i+k,i+\ell;\delta} \tag{4.10}$$

with

$$\begin{aligned}
C_{i-1,j+1} &= \frac{|b_{i-1,j+1}| - b_{i-1,j+1}}{4\Delta x_1\Delta x_2} + \frac{|b_{i,j}| - b_{i,j}}{4\Delta x_1\Delta x_2}\\
C_{i-1,j-1} &= \frac{|b_{i-1,j-1}| + b_{i-1,j-1}}{4\Delta x_1\Delta x_2} + \frac{|b_{i,j}| + b_{i,j}}{4\Delta x_1\Delta x_2}\\
C_{i,j+1} &= \frac{c_{i,j+1} + c_{i,j}}{2\Delta x_2^2} - \frac{|b_{i,j+1}| + |b_{i,j}|}{2\Delta x_1\Delta x_2}\\
C_{i,j-1} &= \frac{c_{i,j-1} + c_{i,j}}{2\Delta x_2^2} - \frac{|b_{i,j-1}| + |b_{i,j}|}{2\Delta x_1\Delta x_2}\\
C_{i+1,j+1} &= \frac{|b_{i+1,j+1}| + b_{i+1,j+1}}{4\Delta x_1\Delta x_2} + \frac{|b_{i,j}| + b_{i,j}}{4\Delta x_1\Delta x_2}\\
C_{i+1,j-1} &= \frac{|b_{i+1,j-1}| - b_{i+1,j-1}}{4\Delta x_1\Delta x_2} + \frac{|b_{i,j}| - b_{i,j}}{4\Delta x_1\Delta x_2}\\
C_{i-1,j} &= \frac{a_{i-1,j} + a_{i,j}}{2\Delta x_1^2} - \frac{|b_{i-1,j}| + |b_{i,j}|}{2\Delta x_1\Delta x_2}\\
C_{i+1,j} &= \frac{a_{i+1,j} + a_{i,j}}{2\Delta x_1^2} - \frac{|b_{i+1,j}| + |b_{i,j}|}{2\Delta x_1\Delta x_2}\\
C_{i,j} &= -\frac{a_{i-1,j} + 2a_{i,j} + a_{i+1,j}}{2\Delta x_1^2}\\
&\quad -\frac{|b_{i-1,j+1}| - b_{i-1,j+1} + |b_{i+1,j+1}| + b_{i+1,j+1}}{4\Delta x_1\Delta x_2}\\
&\quad -\frac{|b_{i-1,j-1}| + b_{i-1,j-1} + |b_{i+1,j-1}| - b_{i+1,j-1}}{4\Delta x_1\Delta x_2}\\
&\quad +\frac{|b_{i-1,j}| + |b_{i+1,j}| + |b_{i,j-1}| + |b_{i,j+1}| + 2|b_{i,j}|}{2\Delta x_1\Delta x_2}\\
&\quad -\frac{c_{i,j-1} + 2c_{i,j} + c_{i,j+1}}{2\Delta x_2^2}
\end{aligned} \tag{4.11}$$

and employ a simple forward difference in time. For this discretisation Weickert has shown that stability in terms of a discrete maximum-minimum principle can only be proven if the spectral condition number of $\mathbf{D}$ is below 5.82. For larger condition numbers he mentioned indications based on experiments that some weaker stability properties might exist.

In conclusion, the discrete analogue of the nonlinear anisotropic diffusion model (4.5) reads in the above terms as

$$\begin{aligned} W^{\star}_{i,j} &= W_{i,j} - \nabla \cdot (\mathbf{D}(W_{i,j}) \nabla W_{i,j}) \\ &= W_{i,j} - \sum_{k=-1}^{1} \sum_{\ell=-1}^{1} C_{i+k,j+\ell} W_{i+k,i+\ell} \end{aligned} \tag{4.12}$$

with $W_{i,j} = U_{i,j;\delta}$ and $U^{i,j} = W^{\star}_{i,j}$.

The resulting algorithm

In the foregoing section we have explained in detail the construction of the operator $\mathcal{C}$ and $\mathcal{D}$, associated with the convective resp. the diffusion step of the algorithm. Combining these parts, we derive a scheme of the form

$$U^{i,j} = \mathcal{D}(\Delta t)\mathcal{C}(\Delta t) U_{i,j}. \tag{4.13}$$

This splitting is known to be first order in time only but the operators described above can also be applied to more sophisticated splittings with second order accuracy like the Strang splitting [103] or the TVD approach by Shu [98]. Nevertheless, in steady state computation like the scalar test case described in the chapter concerned with numerical examples, time accuracy is by no means mandatory. The above coupling can be described in detail in the following manner:

- In every time step:
  1. Compute a numerical solution W by performing the convective step (4.3) on the data u at time level t^n.
  2. If necessary, filter high frequency oscillations by means of discrete convolution with the Gaussian kernel G_σ, i.e. $W_\sigma := G_\sigma * W$.
  3. Compute the structure tensor $\mathbf{J}_0(\nabla W_\sigma)$ which contains information about the local coherence of the numerical solution.
  4. Average the structure information in the vicinity of each grid point according to the integration scale ρ to define a region size in which the orientation of the solution is examined. This corresponds to computing $\mathbf{J}_\rho(\nabla W_\sigma) := G_\rho * \mathbf{J}_0(\nabla W_\sigma)$.
  5. Perform an additional data analysis to gain information on the necessary dissipation strength.
  6. Construct an artificial dissipation tensor $\mathbf{D}$ from the knowledge of the data analysis.
  7. Apply the diffusion step to the data W, i.e. solve the discrete version (4.12) of the nonlinear anisotropic diffusion equation (4.5) with initial data W to derive the filtered data $W^\star$.
- End of time step: Set $U^{n+1} := W^\star$.

Parts of the development of the presented algorithm can be found in [36].

4.2 Data dependent diffusion steering

So far, we have described how to construct a suitable basic discrete dissipation model which includes information about the local orientation and smoothness of the data. This approach possesses the advantage that the algorithm consider important structures in order to adapt the diffusion strength and direction in contrast to a linear isotropic one.

Nevertheless, the Perona-Malik-like diffusion (3.24) still needs several parameters like the threshold value λ and the smoothing scales σ and ρ. On the other hand, it would be useful to get more information from the data. The structure tensor (4.7) reveals information about the orientation of a discontinuity but not about the strength or about the local change of the data. E.g. for a rarefaction wave, the data changes continuously while for shocks and contact discontinuities the data changes significantly from one cell to another.

Therefore it would be useful to incorporate a suitable diffusion model which also takes the strength of the shock better into account. This may be obtained by a special switch or a shock detector. Nevertheless our demand shows clearly the need for an additional analysis of the discrete data.

Coherence measure

One possible way to introduce a kind of shock detection is simply to use the information from the data analysis we have already performed: the structure information gained from the structure tensor (4.7). Since the eigenvalues (3.22) contain information about the local structure of the data or, more precise, the local coherence of the data. Thus we can view the difference of the eigenvalues as a measure of the difference of the data: the so-called coherence measure.

This measure is a very simple and effective tool which results directly from the structure tensor (4.7). It reveals the local structure of the analysed data. We consider again the eigenvalues of the structure tensor (3.22). They describe the contrast of the solution in the directions of the eigenvectors. Since one has

$$\lambda_1 \geq \lambda_2 \geq 0,$$

according to the fact that $\underline{v}_1$ was the eigenvector in the direction of the largest variation, the coherence measure reads in terms of the coefficients of the structure tensor (4.7) as

$$(\lambda_1 - \lambda_2)^2 \quad = \quad (j_{11} - j_{22})^2 + 4j_{12}^2$$

which becomes large for anisotropic structures. Even more, typical features can be characterised in terms of the eigenvalues:

Thus, a possible way to incorporate informations about the anisotropic nature of the data is to weight the diffusion matrix by a measure gained from the coherence. Since the coherence becomes very large for anisotropic structures we shift the coherence data by one and take the logarithm, i.e.

$$\mathrm{coh}(\mathbf{J}_\rho(\nabla U_{i,j;\sigma})) \quad := \quad \ln\left[1 + (\lambda_1 - \lambda_2)^2\right]. \tag{4.14}$$

structures	eigenvalues
constant areas	$\lambda_1 = \lambda_2 = 0$
straight edges	$\lambda_1 \gg \lambda_2 = 0$
corners	$\lambda_1 \geq \lambda_2 \gg 0$

Table 4.1: Accordance between structures and eigenvalues

The result can be seen in Figure 4.2 where a typical flow pattern containing a shock and a rarefaction wave and the corresponding coherence measure (4.14) is presented. The coherence measure computed from the data containing the discontinuity becomes quite large compared to the one for smooth or constant data. Interestingly, even for oscillatory data, e.g. to the left and the right of the shock, the coherence is significantly larger than for areas where the data vary continuously. Thus, the coherence measure (4.14) seems to be a good indicator for discontinuous or oscillatory data.

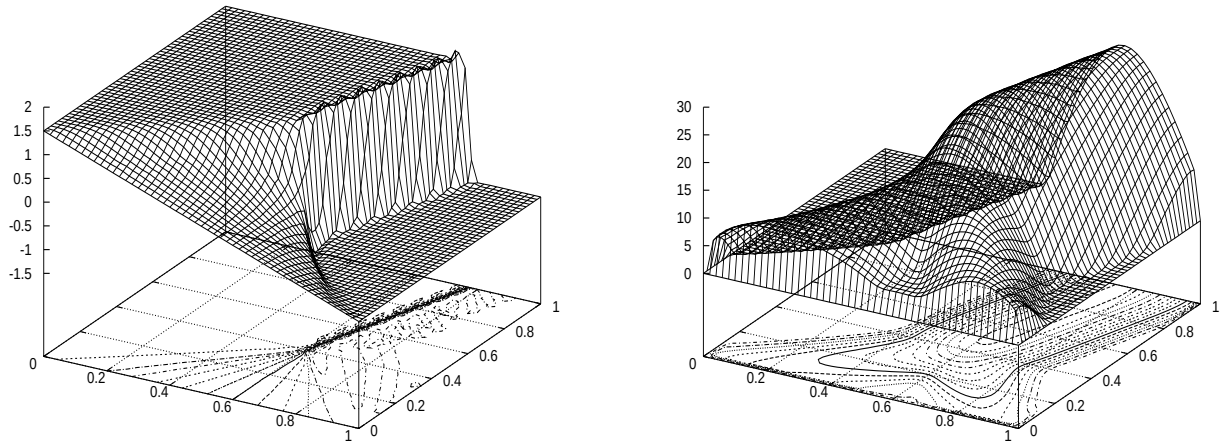

Figure 4.2: Numerical solution and coherence measures for the steady state solution of test problem 6.1.

Another way to exploit this data-dependent steering is the use of a kind of Lax-Wendroff type diffusion steering by weighting the structure tensor with the derivatives of the fluxes. This idea is quite natural if dissipation models of classical finite difference schemes are analysed. Instead of considering the structure tensor (4.7), i.e.

$$\mathbf{J}_\rho(\nabla U) \;=\; \begin{bmatrix} j_{11} & j_{12} \\ j_{12} & j_{22} \end{bmatrix}$$

we employ

$$\tilde{\mathbf{J}}_\rho(\nabla U_{i,j;\sigma}, U_{i,j}) \;=\; \begin{bmatrix} j_{11}(f'(U_{i,j}))^2 & j_{12} f'(U_{i,j}) g'(U_{i,j}) \\ j_{12} f'(U_{i,j}) g'(U_{i,j}) & j_{22}(g'(U_{i,j}))^2 \end{bmatrix}.$$

The resulting scheme is similar to (4.13) in the sense of

$$U^{i,j} = \tilde{\mathcal{D}}(\Delta t)\mathcal{C}(\Delta t)U_{i,j}. \tag{4.15}$$

where $\tilde{\mathcal{D}} = \mathcal{D}(\tilde{\mathbf{J}})$. Results of this approach can be found in [36] and in the chapter concerning the numerical examples.

Shock strength measures

A similar way to detect the shock strength and use it for a diffusion model like (4.12) is to use a 'switch' for the dissipation term in order to determine the required size of the dissipation coefficients. Kreiss and Johansen [60, 40] propose a dissipation model of the form

$$\varepsilon D^+(\phi_i D^- U_i), \tag{4.16}$$

with the shock switch

$$\phi_i = \sum_{k=-p}^{p-1} h|D^+ U_{i+k}|.$$

In regions where the solution is smooth, the dissipation term (4.16) is of order $\mathcal{O}(h^2)$. Thus, if the discretisation of the convective part is second-order accurate it maintains the high order accuracy of the basic scheme. The choice of p is $p = 2$, because efficient methods do not smear discontinuities over more than three grid points.

In the context of the designed filter (4.10) we could define the above switch with the choice $p = 2$ for the corresponding directions as

$$\phi_{i,j}^{x_1} = \sum_{k=-2}^{1} h_1|D_{x_1}^+ U_{i+k,j}| \quad \text{and} \quad \phi_{i,j}^{x_2} = \sum_{k=-2}^{1} h_2|D_{x_2}^+ U_{i,j+k}| \quad .$$

The diffusion matrix (4.6) can be written as

$$\begin{aligned} \tilde{\mathbf{D}}(U_{i,j;\sigma}) &= \begin{bmatrix} \tilde{a} & \tilde{b} \\ \tilde{b} & \tilde{c} \end{bmatrix} \\ &= \begin{bmatrix} (\phi_{i,j;\sigma}^{x_1})^2 a & \phi_{i,j;\sigma}^{x_1}\phi_{i,j;\sigma}^{x_2} b \\ \phi_{i,j;\sigma}^{x_1}\phi_{i,j;\sigma}^{x_2} b & (\phi_{i,j;\sigma}^{x_2})^2 c \end{bmatrix}. \end{aligned}$$

In fact, this idea is somehow related to weighted essentially non-oscillatory (WENO) schemes [76] where weights of the form

$$\begin{aligned} \omega_i^+ &= \left(\frac{1}{\varepsilon + D^+ U_i}\right)^p, \\ \omega_i^- &= \left(\frac{1}{\varepsilon + D^- U_i}\right)^p, \end{aligned}$$

are computed choosing the parameter p as $p = 2$ or $p = 4$. These weights are used to build a reconstruction of the form

$$\tilde{\nabla} U_i \quad := \quad \frac{\omega_i^+ D^+ U_i + \omega_i^- D^- U_i}{\omega_i^+ + \omega_i^-}.$$

This algorithm computes the least oscillatory gradient by a weighted combination of differences of grid values concerning the cell $\mathcal{C}_i$ in one dimension and can be canonically extended to higher dimensions. Of course more sophisticated reconstructions involving more differences, i.e. grid values, are possible.

4.3 Entropy based filter

The concept of involving the coherence measure in the anisotropic diffusion scheme as a shock detector is a very simple and from the computational point of view quite effective approach to steer the diffusion in the vicinity of a shock. On the other hand, it is quite ad hoc even it gives satisfying results. However, since we are concerned with equations arising from some conservation property there are much more powerful tools to describe unsteady solutions, namely the concept of entropy.

We already have mentioned the fact that conservation laws are associated with an entropy – entropy flux pair satisfying an entropy inequality. Moreover, in regions where the solution is smooth we have an entropy equality rather than an inequality. Thus, the theory of entropy-satisfying solutions is not only a very useful tool to distinguish between admissible and physically irrelevant solutions. One may also use it as a detector to distinguish between smooth and discontinuous parts of the solution.

For a numerical scheme one wishes to distinguish between regions which are the vicinity of a shock where we need additional stabilising diffusion, and regions where second-order accurate scheme like central differencing or the Lax-Wendroff scheme can be applied without problems. Using such schemes this task is equivalent to detect areas where unphysical oscillations take place due to instabilities of the numerical scheme near a shock. There the entropy inequality is violated, i.e. is positive and entropy production takes place in this area. Thus, it seems to be a good idea to use the entropy as a reliable descriptor for the existence of an shock curve.

Entropy-steered diffusion

In the following approach, we are going to use the entropy production as a detector for the regions where an additional dissipation model is needed. Thus, we are going to distinguish between regions where the Lax-Wendroff scheme (4.3) produce entropy satisfying solutions and regions where we have to construct an additional dissipation model. This diffusion model will be based on the artificial diffusion filter, constructed in the foregoing section.

The entropy production indicator

Starting from the entropy inequality for the scalar conservation law (4.2), i.e.

$$\partial_t \mathsf{u}(u) + \partial_{x_1} \mathsf{f}(u) + \partial_{x_2} \mathsf{g}(u) \quad \leq \quad 0,$$

the semi-discrete form of this equation is given by

$$\begin{aligned}
&\frac{d}{dt}\mathsf{u}(U_i) + \frac{1}{\Delta x_1}\left[\mathsf{F}_{i+1/2,j} - \mathsf{F}_{i-1/2,j}\right] + \frac{1}{\Delta x_2}\left[\mathsf{G}_{i,j+1/2} - \mathsf{G}_{i,j-1/2}\right] \\
&= \quad \mathsf{u}_i'\left[-\frac{1}{\Delta x_1}\left(F_{i+1/2,j} - F_{i-1/2,j}\right) - \frac{1}{\Delta x_2}\left(G_{i+1/2,j} - G_{i-1/2,j}\right)\right] \\
&\quad + \frac{1}{\Delta x_1}\left[\mathsf{F}_{i+1/2,j} - \mathsf{F}_{i-1/2,j}\right] + \frac{1}{\Delta x_2}\left[\mathsf{G}_{i,j+1/2} - \mathsf{G}_{i,j-1/2}\right] \\
&= \quad \frac{1}{\Delta x_1}\left[\mathsf{F}_{i+1/2,j} - \mathsf{F}_{i-1/2,j} - \mathsf{u}_i'\left(F_{i+1/2,j} - F_{i-1/2,j}\right)\right] \\
&\quad + \frac{1}{\Delta x_2}\left[\mathsf{G}_{i,j+1/2} - \mathsf{G}_{i,j-1/2} - \mathsf{u}_i'\left(G_{i+1/2,j} - G_{i-1/2,j}\right)\right] \;=:\; E_{i,j}.
\end{aligned}$$

Here, the numerical entropy fluxes $\mathsf{F}_{i\pm1/2,i}, \mathsf{G}_{i,j\pm1/2}$ are chosen similar to (2.36) according to the numerical fluxes $F_{i\pm1/2,j}, G_{i,j\pm1/2}$ of the underlying scheme (4.3).

Based on this form we define the numerical entropy production as

$$e_{i,j}^{+} := \begin{cases} E_{i,j} & E_{i,j} > 0 \\ \text{for} & \\ 0 & E_{ij} \leq 0. \end{cases} \tag{4.17}$$

This indicator enables us to design a scheme which distinguishes between regions where the entropy inequality (4.17) is fulfilled and which corresponds to $e_{i,j}^{+} = 0$, and areas where entropy production due to instabilities in the vicinity of discontinuities takes place, i.e. $e_{i,j}^{+} > 0$. This divides the computational domain into areas where the Lax-Wendroff scheme (4.3) can be used, and others where one has to design a suitable dissipation filter like in the foregoing section. In contrast to the diffusion matrix designed there, now we are going to use the gradient of the entropy production.

Remark 4.4
The occurrence of oscillations and the satisfaction of a discrete entropy inequality is related in a highly nontrivial way. E.g. the unmodified Roe scheme [92] possesses the TVD property, but the limit solution allows entropy violating shocks. On the other hand the Lax-Wendroff entropy fix by Majda and Osher [78] produces still oscillations but satisfies a discrete entropy inequality. A hint may be given by the conjecture of Merriam (see [81, 102]). He predicts that a numerical scheme which is stable according to all entropy-entropy flux pairs creates a monotone solution.

Diffusion tensor for smooth solutions

Since a simple central scheme is unconditionally unstable we need a kind of 'background diffusion' that guarantees a stable behaviour of the flow computation. Since in smooth

regions the flow follows the characteristics, we are interested to design a diffusion tensor which reveals this property and gives all dissipation parallel to the characteristic curves, avoiding dissipation perpendicular to them.

Since the characteristic equations for (4.2) are given by $dx_1/ds = f' =: A(u)$ and $dx_2/ds = g' =: B(u)$, we end up with

$$\frac{dx_2}{dx_1} = \frac{B}{A}.$$

The characteristic curves are described locally by the vector $\underline{v} := [A(u),\, B(u)]^T$.

To construct the diffusion tensor $\mathbf{D}_{\mathcal{C}}$ with diffusion direction parallel and perpendicular to the characteristic curves, we take the vectors $[\underline{v}_1, \underline{v}_2] = [\underline{v}, \underline{v}^{\perp}] =: \mathbf{V}$ as the eigenvectors of the tensor and equivalent to (4.9):

$$\mathbf{D}_{\mathcal{C}} \;:=\; \mathbf{V}\mathbf{L}\mathbf{V}^{-1}. \tag{4.18}$$

This reads as

$$\begin{aligned}\mathbf{D}_{\mathcal{C}} \;&:=\; \begin{bmatrix} A & -B \\ B & A \end{bmatrix} \begin{bmatrix} l_1 & 0 \\ 0 & l_2 \end{bmatrix} \frac{1}{A^2+B^2} \begin{bmatrix} A & B \\ -B & A \end{bmatrix} \\ &=\; \frac{1}{A^2+B^2} \begin{bmatrix} l_1A^2 + l_2B^2 & (l_1-l_2)AB \\ (l_1-l_2)AB & l_1B^2 + l_2A^2 \end{bmatrix}.\end{aligned}$$

Since we are interested in restricting diffusion perpendicular to the characteristics we set the eigenvalue corresponding to the vector $\underline{v}_1$ to one, i.e. $l = 1$ and the eigenvalue corresponding to $\underline{v}_2$ to zero, i.e. $l_2 = 0$. Hence, one gets

$$\mathbf{D}_{\mathcal{C}} \;:=\; \frac{1}{A^2+B^2} \begin{bmatrix} A^2 & AB \\ AB & B^2 \end{bmatrix}.$$

For a numerical analyst this looks very familiar and we come to the following

Lemma 4.5
Consider the linearised form of (4.2), i.e.

$$\partial_t u + A\partial_{x_1} u + B\partial_{x_2} u \;=\; 0. \tag{4.19}$$

Than the Lax-Wendroff scheme is the scheme with optimal diffusion in the sense that it induces dissipation only parallel to the characteristic curves.

Proof The Taylor series expansion for u, based on (4.19) reads with $\Delta t = k$ as

$$u(t+k, \underline{x}) \;=\; u(t, \underline{x}) + k u_t(t, \underline{x}) + \tfrac{1}{2}k^2 u_{tt}(t, \underline{x}) + \mathcal{O}(k^3).$$

Plugging in the time derivate $u_t \;=\; -Au_x - Bu_y$ and setting $x_1 = x, x_2 = y$, one gets

$$\begin{aligned} u(t+k, \underline{x}) \;&=\; u(t, \underline{x}) + k(-Au_x - Bu_y) + \tfrac{1}{2}k^2(-Au_x - Bu_y)_t + \mathcal{O}(k^3) \\ &=\; u(t, \underline{x}) - k(Au_x + Bu_y) \\ &\qquad -\tfrac{1}{2}k^2[A(-Au_x - Bu_y)_x + B(-Au_x - Bu_y)_y] + \mathcal{O}(k^3) \\ &=\; u(t, \underline{x}) - k(Au_x + Bu_y) \\ &\qquad +\tfrac{1}{2}k^2[(A^2u_x + ABu_y)_x + (ABu_x + B^2u_y)_y] + \mathcal{O}(k^3) \\ &=\; u(t, \underline{x}) - k(Au_x + Bu_y) + \tfrac{1}{2}k^2 \mathrm{div}[\mathbf{D}_{\mathcal{LW}} \nabla u] + \mathcal{O}(k^3) \end{aligned}$$

with

$$\mathbf{D}_{\mathcal{LW}} = \begin{bmatrix} A^2 & AB \\ AB & B^2. \end{bmatrix}$$

■

Diffusion tensor near discontinuities

Having constructed the diffusion tensor for the smoother regions of the data we look at the region where the entropy inequality is violated, which corresponds to entropy production. Thus, we compute the discrete gradient of the entropy production (4.17):

$$\nabla e^+_{i,j} = \underline{e}_1 := \begin{bmatrix} \dfrac{e^+_{i+1,j} - e^+_{i-1,j}}{2\Delta x_1} \\ \dfrac{e^+_{i,j+1} - e^+_{i,j-1}}{2\Delta x_2} \end{bmatrix}. \tag{4.20}$$

Based on this gradient the construction of the diffusion tensor with $\mathbf{V} := [\underline{e}_1, \underline{e}_1^{\perp}] := [\underline{e}_1, \underline{e}_2]$ reads as

$$\mathbf{D}_{\mathcal{E}} = \mathbf{VLV}^{-1}.$$

This is the diffusion tensor used in the vicinity of shocks. In contrast to (4.9) we now use the gradient of the entropy production (4.20) which is identified with the direction perpendicular to the shock.

The question is how to choose the eigenvalues in $L = \mathrm{diag}(l_1, l_2)$ which control the amount of dissipation by their size. This depends on the application and one has to choose them according to the particular test case

In general, we are interested in diffusion parallel to the gradient of the entropy production. Nevertheless, we need some diffusion across the shock in order to stabilise the numerical scheme. To overcome this problem, the diffusion strength is chosen as $l_1 = 5$ and $l_2 = 5 - 2e^+_{i,j}$ for the test case (6.1).

The resulting blended scheme

After the construction of the diffusion tensors $D_{\mathcal{C}} = D_{\mathcal{LW}}$ and $D_{\mathcal{E}}$ we are able to design a numerical scheme with the following features:

- In smooth region dissipation parallel to the characteristics.
- Near discontinuities blend to a dissipation direction parallel to the important features.
- Avoid smearing and enhance the shock.
- Remove the oscillations in the vicinity of shocks occurring from second order methods.

The blending between the two dissipation tensors is simply based on the amount of the entropy production itself: in regions where the solution is smooth, i.e. $E_{i,j} = 0$, we take the diffusion tensor $\mathbf{D}_{\mathcal{C}}$ and in regions of entropy production blending to $\mathbf{D}_{\mathcal{E}}$ takes place. This is done by the weighting $\phi := 1 - 2 \cdot \nabla e^+_{i,j}$ and $\phi^{'} := 2 \cdot \nabla e^+_{i,j}$.

The resulting scheme writes as

$$\begin{aligned} U^{n+1}_{i,j} &= U^n_{i,j} - \lambda[F_{i+1/2,j} - F_{i-1/2,j}] - \lambda[G_{i,j+1/2} - G_{i,j-1/2}] \\ &\quad + \phi\frac{\lambda}{2}[\mathrm{div}D_{LW}(U_{i,j})\nabla U_{i,j}] + \phi^{'}\frac{\lambda}{2}[\mathrm{div}D_E(U_{i,j})\nabla U_{i,j}]. \end{aligned} \tag{4.21}$$

The discrete divergence operator $\mathrm{div}D(U_{i,j})\nabla U_{i,j}$ is discretised according to (4.10). This algorithm is published in [39].

4.4 Positive filter

Again we start from the indicator for entropy production (4.17) as a basic tool for data analysis in this scheme. It serves as an indicator for regions where one has to use additional methods to correct the scheme in order to reduce unphysical oscillations. The idea behind the following algorithm is the use of the alternative definition of a monotone scheme (2.10), (2.11), which in this case reads

$$U^{i,j} = \sum_{i,j=-1}^{1} c_{i,j}U_{i,j}, \quad c_{i,j} > 0 \quad \forall i,j. \tag{4.22}$$

Thus, if the detector (4.17) indicates instabilities inside the solution we have to rearrange the numerical scheme (4.22) in order to derive positive coefficients which is equivalent to a monotone method. This can be accomplished by an artificial diffusion filter.

The basic scheme

In contrast to the foregoing sections we use a modification of the Lax-Wendroff scheme (4.3) proposed by LeVeque [72, 73]. This scheme uses an alternative discretisation of the cross fluxes at point $(x_{i+1/2}, y_{j+1/2})$ with coefficients AB having $A = f^{'}, B = g^{'}$ and $x = x_1, y = x_2$. Choosing

$$\begin{aligned} A^+ &:= \tfrac{1}{2}(|A| + A), \\ A^- &:= -\tfrac{1}{2}(|A| - A), \\ B^+ &:= \tfrac{1}{2}(|B| + B), \\ B^- &:= -\tfrac{1}{2}(|B| - B). \end{aligned} \tag{4.23}$$

one is able to write the cross flux as

$$AB = \tfrac{1}{4}(A^+B^+ + A^+B^- + A^-B^+ + A^-B^-)$$

and gets

$$\begin{aligned} ABu_y &\approx \tfrac{1}{4h}AB(D_y^- U_{i-1,j} + D_y^- U_{i-1,j+1} + D_y^- U_{i,j} + D_y^- U_{i,j+1}) \\ &\approx \tfrac{1}{4h}(A^+B^+D_y^- U_{i-1,j} + A^+B^-D_y^- U_{i-1,j+1} + A^-B^+D_y^- U_{i,j} + A^-B^-D_y^- U_{i,j+1}). \end{aligned} \tag{4.24}$$

A^+ (resp. A^-) represents a correction wave from the left (resp. right) while B^+ (resp. B^-) represents information travelling upward (resp. downward) into the corresponding cell. This involves a kind of upwind idea into the discretisation of the cross fluxes due to the choice (4.23).

Thus, a cell $\mathcal{C}_{ij}$ is only updated by a diagonally travelling wave if this wave is moving into the cell.

Remark 4.6
For systems of conservation laws one uses the matrix $\mathbf{A} = \mathbf{A}(\underline{U}_{i+1}, \underline{U}_i)$ *computed from the Roe-average [92] and applies the above switch to the eigenvalues, i.e.* $\mathbf{A}^\pm = \mathbf{R}_A \mathbf{\Lambda}_A^\pm \mathbf{R}_A^{-1}$, *where* $\mathbf{R}, \mathbf{\Lambda}$ *are the corresponding matrix of right eigenvectors resp. eigenvalues.*

The numerical scheme reads as

$$U^{i,j} = U_{i,j} - \lambda^x \left[F_{i+1/2,j} - F_{i-1/2,j}\right] - \lambda^y \left[G_{i,j+1/2} - G_{i,j-1/2}\right]. \tag{4.25}$$

The numerical flux with modification (4.24) is given by

$$\begin{aligned} F_{i+1/2,j} &:= \hat{F}_{i+1/2,j} - \tfrac{1}{2}\lambda^x \tilde{F}_{i+1/2,j} \\ \tilde{F}_{i+1/2,j} &:= (A^-B^-)_{i+1,j+1/2} D_y^- U_{i+1,j+1} + (A^+B^-)_{i,j+1/2} D_y^- U_{i,j+1} \\ &\quad + (A^-B^+)_{i+1,j-1/2} D_y^- U_{i+1,j} + (A^+B^+)_{i,j-1/2} D_y^- U_{i,j}. \end{aligned} \tag{4.26}$$

Here, $\hat{F}_{i+1/2,j}$ is the usual Lax-Wendroff flux for the right cell face, i.e.

$$\hat{F}_{i+1/2,j} = \tfrac{1}{2}(F_{i+1} + F_i) - \tfrac{1}{2}\lambda^x A^2_{i+1/2,j}. \tag{4.27}$$

The fluxes $F_{i-1/2,j}, G_{i,j+1/2}$, $G_{i,j-1/2}$ are analogously defined.

The interesting observation concerning this scheme is the following: If we look at the differences of the numerical flux functions we can interpret the discretisation of the cross derivatives as an anisotropic diffusion with nonlinear diffusion coefficients A^+, A^-, B^+, B^-:

$$\begin{aligned} \tilde{F}_{i+1/2,j} - \tilde{F}_{i-1/2,j} &\approx \partial_x (A^-B^- \partial_y U)_{i+1/2,j+1/2} + \partial_x (A^-B^+ \partial_y U)_{i+1/2,j-1/2} \\ &+ \partial_x (A^+B^- \partial_y U)_{i-1/2,j+1/2} + \partial_x (A^+B^+ \partial_y U)_{i-1/2,j-1/2} \end{aligned}$$

$$\begin{aligned} \tilde{G}_{i,j+1/2} - \tilde{G}_{i,j-1/2} &\approx \partial_y (A^-B^- \partial_x U)_{i+1/2,j+1/2} + \partial_y (A^-B^+ \partial_x U)_{i+1/2,j-1/2} \\ &+ \partial_y (A^+B^- \partial_x U)_{i-1/2,j+1/2} + \partial_y (A^+B^+ \partial_x U)_{i-1/2,j-1/2}. \end{aligned}$$

From this modification one derives a dissipation tensor similar to (4.6). Here, the cross fluxes are evaluated at the cell corners and serve as a switch for the diagonally travelling waves, depending whether they go inside or outside of the cell. Hence, the question arises how to modify these corrections to get an oscillation free algorithm. As already mentioned we use the discrete entropy inequality as an indicator, where we have to correct the algorithm in order to get a monotone scheme

Positivity conditions

If the indicator for entropy production is different from zero oscillations occur and the monotonicity condition (4.22) is not fulfilled. Thus, one has to check the coefficients of the scheme. In order to do this we write the discretisation (2.3) with the usual numerical fluxes of the Lax-Wendroff scheme (4.27) as

$$U^{i,j} = \sum_{l,k=-1}^{1} c_{lk} U_{i+l,j+k},$$

with coefficients

$$\begin{aligned}
c_{i,j} &: 1 - \tfrac{1}{2}\lambda^2 \left[A^2_{i+1/2,j} + A^2_{i-1/2,j} + B^2_{i,j+1/2} + B^2_{i,j-1/2} \right], \\
c_{i+1,j} &: \tfrac{1}{2}\lambda \left[\lambda A^2_{i+1/2,j} - A_{i+1,j} \right], \\
c_{i-1,j} &: \tfrac{1}{2}\lambda \left[\lambda A^2_{i-1/2,j} - A_{i-1,j} \right], \\
c_{i,j+1} &: \tfrac{1}{2}\lambda \left[\lambda B^2_{i,j+1/2} - B_{i,j+1} \right], \\
c_{i,j-1} &: \tfrac{1}{2}\lambda \left[\lambda B^2_{i,j-1/2} - B_{i,j-1} \right].
\end{aligned} \tag{4.28}$$

If we are taking into account the discretisation of the cross derivate (4.24) they are modified in the following way:

$$\begin{aligned}
\tilde{c}_{i,j} &:= c_{i,j} + A^-(B^- - B^+)_{i+1/2,j} + A^+(B^+ - B^-)_{i-1/2,j} \\
&\quad + B^-(A^- - A^+)_{i,j+1/2} + B^+(A^+ - A^-)_{i,j-1/2} \\
\tilde{c}_{i+1,j} &:= c_{i+1,j} + A^-(B^+ - B^-)_{i+1/2,j} - (A^-B^-)_{i+1,j+1/2} + (A^-B^+)_{i+1,j-1/2} \\
\tilde{c}_{i-1,j} &:= c_{i-1,j} + A^-(B^+ - B^-)_{i-1/2,j} - (A^+B^+)_{i-1,j-1/2} + (A^+B^-)_{i-1,j+1/2} \\
\tilde{c}_{i,j+1} &:= c_{i,j+1} + B^-(A^+ - A^-)_{i,j+1/2} - (A^-B^-)_{i+1/2,j+1} + (A^+B^-)_{i-1/2,j+1} \\
\tilde{c}_{i,j-1} &:= c_{i,j-1} + B^+(A^- - A^+)_{i,j-1/2} - (A^+B^+)_{i-1/2,j-1} + (A^-B^+)_{i+1/2,j-1} \\
\tilde{c}_{i+1,j+1} &:= c_{i+1,j+1} + (A^-B^-)_{i+1,j+1/2} + (A^-B^-)_{i+1/2,j+1} \\
\tilde{c}_{i-1,j+1} &:= c_{i-1,j+1} - (A^+B^-)_{i-1,j+1/2} - (A^+B^-)_{i-1/2,j+1} \\
\tilde{c}_{i+1,j-1} &:= c_{i+1,j-1} - (A^-B^+)_{i+1,j-1/2} - (A^-B^+)_{i-1/2,j-1} \\
\tilde{c}_{i-1,j-1} &:= c_{i-1,j-1} + (A^+B^+)_{i-1,j+1/2} + (A^+B^+)_{i-1/2,j-1}.
\end{aligned}$$

For the scheme (4.25),(4.26) the notable observation is that the coefficients (4.28) are altered in the following way:

$$\begin{array}{ccccccc}
 & & \tilde{c}_{i,j} & \geq & c_{i,j}, & & \\
\tilde{c}_{i+1,j} & \leq & c_{i+1,j}, & & \tilde{c}_{i+1,j+1} & \geq & 0, \\
\tilde{c}_{i-1,j} & \leq & c_{i-1,j}, & & \tilde{c}_{i+1,j-1} & \geq & 0, \\
\tilde{c}_{i,j+1} & \leq & c_{i,j+1}, & & \tilde{c}_{i-1,j+1} & \geq & 0, \\
\tilde{c}_{i,j-1} & \leq & c_{i,j-1}, & & \tilde{c}_{i-1,j-1} & \geq & 0.
\end{array}$$

Here, one sees that the coefficients concerning the main axes are become less or equal compared to the original Lax-Wendroff scheme while the coefficients for the diagonals are positive.

The idea is to distribute the positive weights from the diagonal axes to the major ones in order to derive a positive scheme. Or, at least, to get the scheme (4.25),(4.26) 'more' positive. Thus, in the following we construct an anisotropic diffusion filter which diffuses from the corners to the major axes to reduce the oscillations at regions where entropy production takes place. As we have seen these regions can be identified as regions where the monotonicity criterion (4.22) is violated, i.e. one of the coefficients $c_{i\pm1,j\pm1}$ is negative.

The discrete positive filter

The diffusion filter is constructed in the following way: If the indicator (4.17) is different from zero, i.e. indicates a region with entropy violating oscillations, the sum of the positive coefficients is computed, i.e.

$$\begin{aligned} \mathcal{C}^+ &= \max(0.0, \tilde{c}_{i+1,j+1}) + \max(0.0, \tilde{c}_{i+1,j-1}) \\ &\quad + \max(0.0, \tilde{c}_{i-1,j+1}) + \max(0.0, \tilde{c}_{i-1,j-1}). \end{aligned} \tag{4.29}$$

The sum of the coefficients which are negative is

$$\begin{aligned} \mathcal{C}^- &= \min(0.0, \tilde{c}_{i,j}) + \min(0.0, \tilde{c}_{i+1,j}) \\ &\quad + \min(0.0, \tilde{c}_{i-1,j}) + \min(0.0, \tilde{c}_{i,j+1}) + \min(0.0, \tilde{c}_{i,j-1}). \end{aligned} \tag{4.30}$$

Since we only distribute the artificial dissipation build in the numerical scheme rather than introduce additional one, the dissipation coefficients are distributed in order to get the scheme positive (or at least 'less' negative).

The amount that we can distribute is characterised by the ratio that we want to distribute, i.e. $\mathcal{C}^-$, and that we can distribute, i.e. $\mathcal{C}^+$. Since we do not want to create new negative weights this ratio has to be limited by unity. Hence the distribution ratio reads as

$$\mathcal{D} := \min(1, \mathcal{C}^-/\mathcal{C}^+), \tag{4.31}$$

and the diffusion coefficients for the cross diffusion weights read as

$$a^{i,j}_{\pm1/2,\pm1/2} := \tilde{c}_{i\pm1,j\pm1}\mathcal{D}. \tag{4.32}$$

The diffusion aligned with the main axes depends on the amount the central coefficient gains from the corners, at least if one does not introduce additional numerical dissipation into the scheme. Thus, we have to compute the amount which we can distribute, without creating new negative weights:

$$\mathcal{C}^+_{i,j} = \sum_{l,k=-1/2}^{1/2} a^{i,j}_{k,l} + \tilde{c}_{i,j}, \tag{4.33}$$

and the sum of the coefficients which are negative,

$$\begin{aligned} \mathcal{C}^-_{i,j} &= \min(0.0, \tilde{c}_{i+1,j}) + \min(0.0, \tilde{c}_{i-1,j}) \\ &\quad + \min(0.0, \tilde{c}_{i,j+1}) + \min(0.0, \tilde{c}_{i,j-1}). \end{aligned} \tag{4.34}$$

The corresponding weights for the diffusion correction reads as

$$\begin{aligned} a^{i,j}_{\pm 1/2,0} &:= |\min(0.0, \tilde{c}_{i\pm 1,j})|\mathcal{D}_{i,j}, \\ a^{i,j}_{0,\pm 1/2} &:= |\min(0.0, \tilde{c}_{i,j\pm 1})|\mathcal{D}_{i,j}, \end{aligned} \tag{4.35}$$

with the central distribution coefficient

$$\mathcal{D}_{i,j} := \min(1, \mathcal{C}^-_{i,j}/\mathcal{C}^+_{i,j}). \tag{4.36}$$

The positive filter

Having computed the above coefficients, we have all essential ingredients we need for the positive filter. All we have to do is to compute the discretisation of the diffusion tensor $\mathbf{D}$, similar to (4.6). Computing these coefficients for each cell we have to take the coefficients from neighbouring fluxes into account in order to derive a conservative scheme. A possible procedure is displayed in the following compact summary of the algorithm.

The algorithm

- for each cell:
  - if numerical entropy production takes place (i.e. $e^+_{i,j} > 0$):
    1. compute the diffusion coefficients (4.32) for the diagonal dissipation from the distribution coefficient $\mathcal{D}$ (4.31). This is the limited ratio of the sum of the positive and negative weights $\mathcal{C}^+, \mathcal{C}^-$, i.e. (4.29),(4.30).
    2. compute the diffusion coefficients (4.35) concerning the main axes from the central distribution coefficient (4.36). This is the limited ratio of the sum of the central positive and central negative weights $\mathcal{C}^+, \mathcal{C}^-$, i.e. (4.33),(4.34).
- compute the diffusion coefficients $d_{i\pm 1/2,j}, d_{i,j\pm 1/2}, d_{i\pm 1/2,j\pm 1/2}$ as functions of the corresponding diffusion coefficients (4.32),(4.35) of the neighbouring cells:

$$\begin{aligned} d_{i+1/2,j} &:= \mathrm{m}(a^{i,j}_{+1/2,0}, a^{i+1,j}_{-1/2,0}), \\ d_{i-1/2,j} &:= \mathrm{m}(a^{i,j}_{-1/2,0}, a^{i-1,j}_{+1/2,0}), \\ d_{i,j+1/2} &:= \mathrm{m}(a^{i,j}_{0,+1/2}, a^{i,j+1}_{0,-1/2}), \\ d_{i,j-1/2} &:= \mathrm{m}(a^{i,j}_{0,-1/2}, a^{i,j-1}_{0,+1/2}), \\ d_{i+1/2,j+1/2} &:= \mathrm{m}(a^{i,j}_{+1/2,+1/2}, a^{i+1,j+1}_{-1/2,-1/2}), \\ d_{i+1/2,j-1/2} &:= \mathrm{m}(a^{i,j}_{+1/2,-1/2}, a^{i+1,j-1}_{-1/2,+1/2}), \\ d_{i-1/2,j-1/2} &:= \mathrm{m}(a^{i,j}_{-1/2,-1/2}, a^{i-1,j-1}_{+1/2,+1/2}), \\ d_{i-1/2,j+1/2} &:= \mathrm{m}(a^{i,j}_{-1/2,+1/2}, a^{i-1,j+1}_{+1/2,-1/2}). \end{aligned}$$

Here, m means taking a function like the maximum, minimum or arithmetical average of the arguments. The maximum may lead to instabilities while the minimum works in every case.

With the above notation, the resulting scheme can be written as

$$\begin{aligned} U^{i,j} &= \hat{U}_{i,j} + \tfrac{1}{2}\lambda^x \left[d_{i+1/2,j}(U_{i+1,j} - U_{i,j}) - d_{i-1/2,j}(U_{i,j} - U_{i-1,j})\right] \\ &\quad + \tfrac{1}{2}\lambda^y \left[d_{i,j+1/2}(U_{i,j+1} - U_{i,j}) - d_{i,j-1/2}(U_{i,j} - U_{i,j-1})\right] \\ &\quad + \tfrac{1}{2}\lambda^{xy}\big[d_{i+1/2,j+1/2}(U_{i+1,j+1} - U_{i,j}) - d_{i-1/2,j-1/2}(U_{i,j} - U_{i-1,j-1}) \\ &\quad\quad + d_{i-1/2,j+1/2}(U_{i-1,j+1} - U_{i,j}) - d_{i+1/2,j-1/2}(U_{i,j} - U_{i+1,j-1})\big] \end{aligned} \tag{4.37}$$

with

$$\hat{U}_{i,j} = U^{i,j} - \lambda^x \left[F_{i+1/2,j} - F_{i-1/2,j}\right] - \lambda^y \left[G_{i,j+1/2} - G_{i,j-1/2}\right]$$

and $\lambda^{xy} = \Delta t/\sqrt{\Delta x^2 + \Delta y^2}$. It is possible to write this in a more compact way, but to distinguish here between the filter and the basic step we use this extended notation.

Remark 4.7
(4.37) can be regarded as the LeVeque scheme (4.25) enriched with the discretisation of a diffusion filter of the form (4.6). The resulting scheme again is anisotropic and data-dependent. In contrast, here we do not need an additional positive discretisation like (4.11). This is already build in the construction, e.g. with the coefficients A^+, A^-, B^+, B^-.

It is necessary to state that since the discrete filter is data-dependent it will not be monotone in every case. This is due to the fact, that no additional numerical dissipation is added but only the existing diffusion distributed in an optimal way. So situations may occur, where it is not possible to make all coefficients of (4.28) positive. Due to this fact we consider this scheme as being *quasi monotone*. The construction of this algorithm were already presented [38].

The need for high-resolution schemes is a direct consequence of the nonlinear properties of systems of hyperbolic conservation laws such as the Euler equations.

B. van Leer ("Upwind and High-Resolution Schemes" [54])

5 Discrete filters for the Euler equations

So far we have designed filters for scalar conservation laws. Naturally the question arises how anisotropic diffusion models can be applied to systems of conservation laws. Since the Euler equations are one of the most prominent and important systems in the theory of conservation laws it is reasonable to construct filter algorithms for these equations.

It is clear that the construction of such schemes is by far more difficult than in the scalar case. The strong nonlinearity of the equations due to the coupling of the equations – namely the equations for the conservation of mass, momentum and energy – leads to several problems in the design of the algorithm. Nevertheless, the encouraging results for scalar problems let us break this new ground.

Based on the approach presented in [34, 35] we are going to design a nonlinear anisotropic diffusion filter for the Euler equation which is accomplished by a characteristic decoupling.

5.1 The Euler equations

We consider the two dimensional time-dependent Euler equations in conserved form, i.e.

$$\partial_t \underline{u} + \partial_{x_1} \underline{f}_1(\underline{u}) + \partial_{x_2} \underline{f}_2(\underline{u}) \quad = \quad 0. \tag{5.1}$$

This is a system of nonlinear hyperbolic conservation laws that governs the dynamics of a compressible material such as gases or liquids at high pressure for which the effects of body forces, viscous stresses, and heat flux are neglected.

In the choice of the set of variables, there is some degree of freedom to describe the flow. One distinguishes between primitive variables (or physical variables) which are the density ρ, pressure p and velocity $\underline{v} = (v_1, v_2)^T$ and the conserved variables, which result directly from the fundamental laws of conservation, i.e. conservation of mass, momentum and energy and leads to the conservative formulation (5.1).

We already have demonstrated in the second chapter that there are some advantages in expressing the governing equations (5.1) in this formulation, e.g. results concerning convergence and consistency.

Conserved Variables

For the Euler equations in conserved form, i.e. (5.1), the vectors

$$\underline{u} = \begin{bmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho E \end{bmatrix} \quad , \quad \underline{f}_1(\underline{u}) = \begin{bmatrix} \rho v_1 \\ \rho v_2^2 + p \\ \rho v_1 v_2 \\ \rho H v_1 \end{bmatrix} \quad , \quad \underline{f}_2(\underline{u}) = \begin{bmatrix} \rho v_1 \\ \rho v_2 v_1 \\ \rho v_2^2 + p \\ \rho H v_2 \end{bmatrix}, \tag{5.2}$$

describe the vector of conserved quantities and the flux function in x_1 and x_2, respectively. Since $t \in \mathbb{R}_0^+ := \{t \in \mathbb{R} \,|\, t \geq 0\}$ indicates time and $\underline{x} = (x_1, x_2)^T \in \mathbb{R}^2$ space coordinates the mapping

$$\mathbb{R}_0^+ \times \mathbb{R}^2 \ni (t, \underline{x}) \overset{\rho,\underline{v},p,E,H}{\mapsto} (\rho, \underline{v} := (v_1, v_2)^T, p, E, H)(t, \underline{x})$$

denotes density, velocity, pressure, total energy and enthalpy, respectively. Enthalpy is defined by

$$H \quad := \quad E + \frac{p}{\rho}.$$

To close the system an equation of state is needed. For ideal gases one uses

$$p = (\gamma - 1)\rho \left(E - \frac{|v|^2}{2} \right)$$

where γ denotes the ratio of specific heats. In the case of dry air one assumes a value of $\gamma = 1.4$.

The Jacobians of the flux functions $\underline{f}_1 = \underline{f}$ and $\underline{f}_2 = \underline{g}$ are given by

$$\mathbf{A} \quad := \quad \nabla_{\underline{u}} f(\underline{u}) \; = \; \begin{bmatrix} 0 & 1 & 0 & 1 \\ \frac{\gamma-3}{2} v_1^2 + \frac{\gamma-1}{2} v_2^2 & (3-\gamma) v_1 & (1-\gamma) v_2 & \gamma - 1 \\ -v_1 v_2 & v_2 & v_1 & 0 \\ (\gamma-1) v_1 |v|^2 - \gamma v_1 E & \gamma E - \frac{\gamma-1}{2}(v_2^2 + 3 v_1^2) & (1-\gamma) v_1 v_2 & \gamma v_1 \end{bmatrix},$$

$$\mathbf{B} \quad := \quad \nabla_{\underline{u}} g(\underline{u}) \; = \; \begin{bmatrix} 0 & 0 & 1 & 1 \\ -v_1 v_2 & v_2 & v_1 & 0 \\ \frac{\gamma-3}{2} v_2^2 + \frac{\gamma-1}{2} v_1^2 & (1-\gamma) v_1 & (3-\gamma) v_2 & \gamma - 1 \\ (\gamma-1) v_2 |v|^2 - \gamma v_2 E & (1-\gamma) v_1 v_2 & \gamma E - \frac{\gamma-1}{2}(v_1^2 + 3 v_2^2) & \gamma v_2 \end{bmatrix},$$

respectively. A thorough discussion concerning theory as well as applications and computational aspects for the Euler equations can be found in the textbooks of Hirsch [51, 52].

Characteristic decomposition

If we assume for a moment that the quantities describing (5.1) are sufficiently smooth we can write the Euler equations in quasilinear form, i.e.

$$\partial_t \underline{u} + \mathbf{A} \partial_{x_1} \underline{u} + \mathbf{B} \partial_{x_2} \underline{u} \quad = \quad 0, \tag{5.3}$$

with matrices $\mathbf{A}, \mathbf{B}$ described above. If we consider a matrix $\mathbf{C}$ of the form (1.28), i.e.

$$\mathbf{C}(\underline{u}, \underline{n}) \;:=\; \mathbf{A}n_{x_1} + \mathbf{B}n_{x_2}$$

for an arbitrary unit vector $\underline{n} = (n_{x_1}, n_{x_2})^T$ one can write the decomposition of $\mathbf{C}$ as

$$\mathbf{C} \;=\; \mathbf{R}\mathbf{\Lambda}\mathbf{R}^{-1}.$$

$\mathbf{R}$ is the matrix of right eigenvectors of $\mathbf{C}$ which read as

$$\underline{r}_1 = \begin{bmatrix} 1 \\ v_1 \\ v_2 \\ \frac{1}{2}|\underline{v}|^2 \end{bmatrix}, \quad \underline{r}_2 = \begin{bmatrix} 0 \\ \rho n_{x_2} \\ -\rho n_{x_2} \\ \rho(v_1 n_{x_2} - v_2 n_{x_1}) \end{bmatrix}, \quad \underline{r}_3 = \frac{\rho}{2c} \begin{bmatrix} 1 \\ (v_1 + c n_{x_1}) \\ (v_2 + c n_{x_2}) \\ H + c\langle \underline{v}, \underline{n} \rangle \end{bmatrix}, \quad \underline{r}_4 = \frac{\rho}{2c} \begin{bmatrix} 1 \\ (v_1 - c n_{x_1}) \\ (v_2 - c n_{x_2}) \\ H - c\langle \underline{v}, \underline{n} \rangle \end{bmatrix},$$

and $\mathbf{\Lambda}$ is the diagonal matrix

$$\mathbf{\Lambda} = \begin{bmatrix} \langle \underline{v}, \underline{n} \rangle & & & 0 \\ & \langle \underline{v}, \underline{n} \rangle & & \\ & & \langle \underline{v}, \underline{n} \rangle + c & \\ 0 & & & \langle \underline{v}, \underline{n} \rangle - c \end{bmatrix}.$$

Thus, the matrix $\mathbf{R}$ results for $\underline{n} = (1, 0)^T$ in the matrix of right eigenvectors for $\mathbf{A}$, respective for $\underline{n} = (0, 1)^T$ in the matrix for $\mathbf{B}$. The same holds obviously for $\mathbf{\Lambda}$.

5.2 Characteristic filters

The use of numerical methods for the Euler equations based on the characteristic decomposition of the linearised form (5.3) is widely used in the area of computational fluid dynamics. A necessary demand in this context is that the numerical approximation for the Jacobian $\mathbf{A}$ of a flux function $\underline{f}$ fulfils some requirements like

$$\underline{f}(\underline{U}_r) - \underline{f}(\underline{U}_l) \;=\; \mathbf{A}(\underline{U}_l, \underline{U}_r)(\underline{U}_r - \underline{U}_l),$$

which were formulated by Roe and is accomplished by evaluating $\mathbf{A}(\underline{U}_l, \underline{U}_r)$ by the Roe-averaged quantities [92]. Some standard approaches make use of upwind-ideas, i.e. a splitting according to the sign of the eigenvalues and flux limiters which act on the characteristic variables. We present here the application of anisotropic diffusion filters developed previously for the scalar case to the Euler equations by characteristic formulation.

Construction of the characteristic filter

We start from an explicit central scheme written as

$$\underline{U}^{i,j} = \underline{U}_{i,j} - \Delta t \left(\frac{1}{\Delta x_1} [\tilde{\underline{F}}_{i+1/2,j} - \tilde{\underline{F}}_{i-1/2,j}] + \frac{1}{\Delta x_2} [\tilde{\underline{G}}_{i,j+1/2} - \tilde{\underline{G}}_{i,j-1/2}] \right).$$

The numerical flux functions include the filter terms and hence the cross diffusion parts and can be written as follows:

$$\begin{aligned}
\tilde{\underline{F}}_{i+1/2,j} &= [\underline{F}_{i+1/2,j} + \underline{L}^{\mathbf{A}}_{i+1/2,j} + \underline{L}^{\mathbf{C}^-}_{i+1/2,j-1/2}],\\
\tilde{\underline{F}}_{i-1/2,j} &= [\underline{F}_{i-1/2,j} + \underline{L}^{\mathbf{A}}_{i-1/2,j} + \underline{L}^{\mathbf{C}^-}_{i-1/2,j+1/2}],\\
\tilde{\underline{G}}_{i,j+1/2} &= [\underline{G}_{i,j+1/2} + \underline{L}^{\mathbf{B}}_{i,j+1/2} + \underline{L}^{\mathbf{C}^+}_{i+1/2,j+1/2}],\\
\tilde{\underline{G}}_{i,j-1/2} &= [\underline{G}_{i,j-1/2} + \underline{L}^{\mathbf{B}}_{i,j-1/2} + \underline{L}^{\mathbf{C}^+}_{i-1/2,j-1/2}].
\end{aligned}$$

The filter operators $\underline{L}^{\mathbf{A}}, \underline{L}^{\mathbf{B}}$ depending on the Jacobian-matrices $\mathbf{A}$ and $\mathbf{B}$ of the flux function $\underline{f}$ and $\underline{g}$, respectively. The filter terms $\underline{L}^{\mathbf{C}^{(\cdot)}}$ models some kind of cross diffusion and depends on a combination of the Jacobians, namely

$$\begin{aligned}
\mathbf{C}^+ &= \mathbf{A}n^+_{x_1} + \mathbf{B}n^+_{x_2},\\
\mathbf{C}^- &= \mathbf{A}n^-_{x_1} + \mathbf{B}n^-_{x_2},
\end{aligned}$$

(see [52] for details). The vector

$$\underline{n}^{(\cdot)} \;:=\; \begin{bmatrix} n^{(\cdot)}_{x_1} \\ n^{(\cdot)}_{x_2} \end{bmatrix}$$

is chosen in the direction of the cell diagonals, i.e

$$\underline{n}^{+} := \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix} \quad \text{and} \quad \underline{n}^{-} := \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}.$$

The matrices all are evaluated as Roe averages (see [92]). Hence one writes the corresponding filter terms as

$$\begin{aligned}
\underline{L}^{\mathbf{A}}_{i+1/2,j} &= \mathbf{R}^{\mathbf{A}}_{i+1/2,j}\underline{\Phi}^{\mathbf{A}}_{i+1/2,j},\\
\underline{L}^{\mathbf{B}}_{i+1/2,j} &= \mathbf{R}^{\mathbf{B}}_{i+1/2,j}\underline{\Phi}^{\mathbf{B}}_{i+1/2,j},\\
\underline{L}^{\mathbf{C}^+}_{i+1/2,j} &= \mathbf{R}^{\mathbf{C}^+}_{i+1/2,j}\underline{\Phi}^{\mathbf{C}^+}_{i+1/2,j},
\end{aligned} \tag{5.4}$$

and similar for the other terms. $\mathbf{R}^{(\cdot)}$ denotes the matrix of the right eigenvectors corresponding to the appropriate matrices written as superscript. $\underline{\Phi}^{(\cdot)}$ denotes the actual filter term. In the basic construction, we follow the recent paper of Yee and her co-workers [118], based on the Artificial Compression Method (ACM) of Harten and the extensions of Yee (see [41, 42, 117]). We extend these method by the integration of a directional based approach: the anisotropic diffusion filter.

The elements of $\underline{\Phi}^{(\cdot)}$ in (5.4) are denoted by $\phi^{l(\cdot)}, l = 1, 2, 3, 4$. Here we state as an example the construction of the term $\phi^{l(\cdot)}_{i+1/2,j}$. All other terms are constructed analogously depending on the matrices and grid points were they are evaluated. Thus, the elements of $\underline{\Phi}^{(\cdot)}_{i+1/2,j}$ read as

$$\underline{\Phi}^{l(\cdot)}_{i+1/2,j} = \beta^l_{i+1/2,j}\phi^l_{i+1/2,j}.$$

β^l denotes the weighting coefficient steering the anisotropic diffusion. The construction of this coefficient will be described in the following section. The choice of the $\phi^l_{i+1/2,j}$ is exactly the same as the one described in [118] with the use of the limiter functions g^l_j like

$$g^l_j = (\alpha^l_{i+1/2,j}\alpha^l_{i-1/2,j} + |\alpha^l_{i+1/2,j}\alpha^l_{i-1/2,j}|)/(\alpha^l_{i+1/2,j} + \alpha^l_{i-1/2,j}).$$

Additional possible choices for this function and a detailed description are given in [118]. The definition of ϕ^l will be given below.

The choice of the filter function depends on the characteristic splitting of the flux representation of the gradients. Thus, we consider $\alpha^l_{i+1/2,j}$ as elements of

$$\underline{\alpha}_{i+1/2,j} := (\mathbf{R}^{\mathbf{A}}_{i+1/2,j})^{-1}(\underline{U}_{i+1,j} - \underline{U}_{i,j}).$$

The different α depend again on different matrices of the right eigenvectors:

$$\begin{aligned}
\underline{\alpha}_{i+1/2,j} &:= (\mathbf{R}^{\mathbf{A}}{}_{i+1/2,j})^{-1}(\underline{U}_{i+1,j} - \underline{U}_{i,j}),\\
\underline{\alpha}_{i-1/2,j} &:= (\mathbf{R}^{\mathbf{A}}{}_{i-1/2,j})^{-1}(\underline{U}_{i,j} - \underline{U}_{i-1,j}),\\
\underline{\alpha}_{i,j+1/2} &:= (\mathbf{R}^{\mathbf{B}}{}_{i,j+1/2})^{-1}(\underline{U}_{i,j+1} - \underline{U}_{i,j}),\\
\underline{\alpha}_{i,j-1/2} &:= (\mathbf{R}^{\mathbf{B}}{}_{i,j-1/2})^{-1}(\underline{U}_{i,j} - \underline{U}_{i,j-1}),\\
\underline{\alpha}_{i+1/2,j+1/2} &:= (\mathbf{R}^{\mathbf{C}^+}{}_{i+1/2,j+1/2})^{-1}(\underline{U}_{i+1,j+1} - \underline{U}_{i,j}),\\
\underline{\alpha}_{i-1/2,j-1/2} &:= (\mathbf{R}^{\mathbf{C}^+}{}_{i-1/2,j-1/2})^{-1}(\underline{U}_{i,j} - \underline{U}_{i-1,j-1}),\\
\underline{\alpha}_{i+1/2,j-1/2} &:= (\mathbf{R}^{\mathbf{C}^-}{}_{i+1/2,j-1/2})^{-1}(\underline{U}_{i+1,j-1} - \underline{U}_{i,j}),\\
\underline{\alpha}_{i-1/2,j+1/2} &:= (\mathbf{R}^{\mathbf{C}^-}{}_{i-1/2,j+1/2})^{-1}(\underline{U}_{i,j} - \underline{U}_{i-1,j+1}).
\end{aligned} \tag{5.5}$$

Hence $\phi^l_{i+1/2,j+1/2}$ corresponds to $\alpha^l_{i+1/2,j+1/2}$ and so on.

The filter function writes as

$$\phi^l_{i+1/2,j} = \frac{1}{2}\psi(a^l_{i+1/2,j})(g^l_{i+1,j} + g^l_{i,j}) - \psi(a^l_{i+1/2,j} + \gamma^l_{i+1/2,j})\alpha^l_{i+1/2,j} \tag{5.6}$$

with

$$\gamma^l_{i+1/2,j} = \frac{1}{2}\psi(a^l_{i+1/2,j}) \begin{cases} (g^l_{i+1,j} - g^l_{i,j})/\alpha^l_{i+1/2,j} & \alpha^l_{i+1/2,j} \neq 0 \\ 0 & \alpha^l_{i+1/2,j} = 0 \end{cases}. \tag{5.7}$$

The $a^l_{i+1/2,j}$ in (5.6),(5.7) are the characteristic speeds of the corresponding Jacobians which are equal to their eigenvalues. This term can also be weighted with the diffusion coefficient $\beta^l_{i+1/2,j}$ steering the amount of dissipation corresponding to this direction. The function ψ is the modulus of the argument, i.e. $\psi(\alpha^l_{i+1/2}) = |\alpha^l_{i+1/2}|$.

The structure tensor

We start from the construction of the structure tensor $\mathbf{J}^0_{i,j}$ similar to the scalar case (4.7). We use a smoothed version of the characteristic gradients $\mathbf{R}^{-1}\nabla\underline{U}_{i,j;\delta}$, where $\underline{U}_{i,j;\delta}$ is a pre-smoothed version of the data $\underline{U}_{i,j}$. This means component-wise convolution with a Gaussian

kernel with convolution scale δ. In the continuous case this is equivalent to solve the heat equation. Thus, we apply a discrete form of the heat equation to the data $U_{i,j}$ with stopping time $T = \frac{1}{2}\delta^2$. In order to remove the small scale oscillations by means of this smoothing technique we define δ depending on the grid size $h := \sqrt{\Delta x_1 \Delta x_2}$, e.g. $\delta = 2h$.

Consequently, the structure tensor reads as

$$\mathbf{J}^0_{i,j} = \begin{pmatrix} j_{11} & j_{12} \\ j_{21} & j_{22} \end{pmatrix} \tag{5.8}$$

with

$$\begin{aligned} j_{11} &:= [0.5(\alpha^{l;\delta}_{i+1/2,j} + \alpha^{l;\delta}_{i-1/2,j})]^2 \\ j_{22} &:= [0.5(\alpha^{l;\delta}_{i,j+1/2} + \alpha^{l;\delta}_{i,j-1/2})]^2 \\ j_{12} &:= j_{11} j_{22} \\ j_{21} &:= j_{12} \end{aligned}$$

where $\alpha^{l;\delta}$ corresponds to the smoothed data $\underline{U}_\delta$. Thereby one can also use a smoothed version of the structure tensor (5.8), i.e

$$\mathbf{J}^\nu_{i,j} := G_\nu * \mathbf{J}^0(\alpha_{i,j}) \tag{5.9}$$

which means componentwise convolution with scale ν, which denotes the width of the averaging region. In practice we are solving the heat equation for each component separately.

The diffusion matrix

After having built the structure tensor we are going to compute the eigenvectors and eigenvalues of (5.9),

$$\underline{v}_{1;\nu} = \begin{pmatrix} 2j_{12} \\ j_{22} - j_{11} + \sqrt{(j_{11} - j_{22})^2 + 4j_{12}^2} \end{pmatrix}, \tag{5.10}$$

$$\underline{v}_{2;\nu} = \begin{pmatrix} j_{11} - j_{22} - \sqrt{(j_{11} - j_{22})^2 + 4j_{12}^2} \\ 2j_{12} \end{pmatrix}. \tag{5.11}$$

Thereby the corresponding eigenvalues are again given by

$$\lambda_{1,2;\nu} = \frac{1}{2}\left(j_{11} + j_{22} \pm \sqrt{(j_{11} - j_{22})^2 + 4j_{12}^2}\right).$$

Now we construct the dissipation matrix $\mathbf{D}$ by introducing the Ansatz due to Weickert [116]:

$$\mathbf{D} := \mathbf{V}_\nu \mathbf{\Lambda}_\nu \mathbf{V}_\nu^{-1}. \tag{5.12}$$

Here $\mathbf{V}_\nu$ denotes the matrix of the eigenvectors (5.10), (5.11) and $\mathbf{\Lambda}_\nu = \mathrm{diag}(l_1, l_2)$ represents a diagonal matrix where the diagonal elements have to be calculated in a convenient manner

as described below. In order to recover shocks (or, equivalently, in order to enhance edges) the diffusivity l_1 perpendicular to edges should be reduced if the contrast $\lambda_{1;\nu}$ is high. This can be achieved by a choice for $\mathbf{\Lambda}$ similar to (3.24).

Thus, the diffusion matrix (5.12) can be written as

$$\mathbf{D} = \begin{bmatrix} a^l & b^l \\ b^l & c^l \end{bmatrix} \tag{5.13}$$

with coefficients

$$\begin{aligned} a^l &= l_1 v_{1_1;\nu}^2 + l_2 v_{1_2;\nu}^2, \\ b^l &= (l_1 - l_2) v_{1_1;\nu} v_{1_2;\nu}, \\ c^l &= l_2 v_{1_1;\nu}^2 + l_1 v_{1_2;\nu}^2, \end{aligned}$$

where $\underline{v}_{1;\rho} = (v_{1_1;\rho}, v_{1_2;\rho})^T$. The superscript l means that the diffusion matrix (5.12) and so the structure tensor (5.8) resp. (5.9) have to be computed for every characteristic variable separately.

The discretisation of the 3×3 stencil reads in accordance with (4.11) as:

$$\begin{aligned} \beta^l_{i-1,j+1} &= \frac{|b_{i-1,j+1}| - b_{i-1,j+1}}{4\Delta x \Delta y} + \frac{|b_{i,j}| - b_{i,j}}{4\Delta x \Delta y}, \\ \beta^l_{i-1,j-1} &= \frac{|b_{i-1,j-1}| + b_{i-1,j-1}}{4\Delta x \Delta y} + \frac{|b_{i,j}| + b_{i,j}}{4\Delta x \Delta y}, \\ \beta^l_{i,j+1} &= \frac{c_{i,j+1} + c_{i,j}}{2\Delta y^2} - \frac{|b_{i,j+1}| + |b_{i,j}|}{2\Delta x \Delta y}, \\ \beta^l_{i,j-1} &= \frac{c_{i,j-1} + c_{i,j}}{2\Delta y^2} - \frac{|b_{i,j-1}| + |b_{i,j}|}{2\Delta x \Delta y}, \\ \beta^l_{i+1,j+1} &= \frac{|b_{i+1,j+1}| + b_{i+1,j+1}}{4\Delta x \Delta y} + \frac{|b_{i,j}| + b_{i,j}}{4\Delta x \Delta y}, \\ \beta^l_{i+1,j-1} &= \frac{|b_{i+1,j-1}| - b_{i+1,j-1}}{4\Delta x \Delta y} + \frac{|b_{i,j}| - b_{i,j}}{4\Delta x \Delta y}, \\ \beta^l_{i-1,j} &= \frac{a_{i-1,j} + a_{i,j}}{2\Delta x^2} - \frac{|b_{i-1,j}| + |b_{i,j}|}{2\Delta x \Delta y}, \\ \beta^l_{i+1,j} &= \frac{a_{i+1,j} + a_{i,j}}{2\Delta x^2} - \frac{|b_{i+1,j}| + |b_{i,j}|}{2\Delta x \Delta y}, \end{aligned}$$

$$\begin{aligned} \beta^l_{i,j} = & -\frac{a_{i-1,j} + 2a_{i,j} + a_{i+1,j}}{2\Delta x^2} \\ & - \frac{|b_{i-1,j+1}| - b_{i-1,j+1} + |b_{i+1,j+1}| + b_{i+1,j+1}}{4\Delta x \Delta y} \\ & - \frac{|b_{i-1,j-1}| + b_{i-1,j-1} + |b_{i+1,j-1}| - b_{i+1,j-1}}{4\Delta x \Delta y} \\ & + \frac{|b_{i-1,j}| + |b_{i+1,j}| + |b_{i,j-1}| + |b_{i,j+1}| + 2|b_{i,j}|}{2\Delta x \Delta y} \\ & - \frac{c_{i,j-1} + 2c_{i,j} + c_{i,j+1}}{2\Delta y^2}. \end{aligned}$$

This gives the weighting coefficients for the dissipative fluxes which leads to a steering of the dissipation terms depending on the magnitude of the gradients as described in the last section.

The numerical results concerning these characteristic filters are given in the next section which is devoted to numerical examples.

I cannot do it without comp[u]ters.

William Shakespeare
("The Winter's Tale")

6 Numerical examples

In this chapter we present the numerical results concerning the schemes developed in the two foregoing sections. First we consider the scalar case. In the second section results for the Euler equations are presented.

6.1 Scalar test case

In this section we present numerical results of the developed schemes, namely the

- basic scheme with coherence measure (4.13),
- basic scheme with weighted coherence measure (4.15),
- entropy steered diffusion (4.21),
- positive entropy filter (4.37).

All schemes are applied to the same scalar test case which includes a shock as well as a fan-like structure. The test case is given by the initial boundary value problem with data

$$u(x,y,0) = \begin{cases} 1.5 & ; \quad x = 0 \\ -2.5x + 1.5 & ; \quad y = 0 \\ -1.0 & ; \quad x = 1 \\ 0 & ; \quad \text{else} \end{cases} \tag{6.1}$$

and fluxes

$$f(u) = 0.5u^2, \qquad g(u) = u.$$

The values at the lateral and lower boundaries are kept fixed. We compute the solution on a Cartesian grid with 50×50 points and determine the boundary condition on the upper side of the unit square by simple extrapolation, then we get a steady solution as shown in Figure 6.1.

The true solution consists of a fan-like continuous wave which develops into a shock. A schematic view of it can be seen in Figure 6.2.

Since the true solution satisfies the equation for the steady state,

$$\partial_y u + \partial_x \frac{u^2}{2} = 0,$$

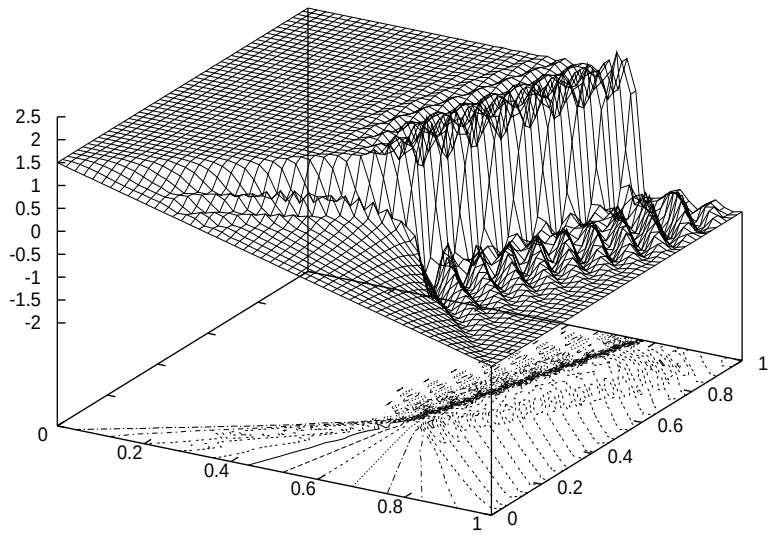

Figure 6.1: Lax-Wendroff solution of test problem

the characteristic equations are given by $dy/ds = 1, dx/ds = u$, i.e.

$$\frac{dy}{dx} = \frac{1}{u}.$$

If we denote by u_L and u_R the given left and right state at $y = 0$, respectively, and we assume a linear distribution

$$u(x, 0) = (u_R - u_L)x + u_L$$

of the boundary data at $y = 0$, then the equation of the leftmost characteristic g_1 is given by $y = x/u_L$. The rightmost characteristic g_2 is given by $y = (x - 1)/u_R$. They meet at the point P where the shock g_3 starts. The coordinates of P are easily computed to be $x_P = u_L/(u_L - u_R)$ and $y_P = 1/(u_L - u_R)$. From the Rankine-Hugoniot condition we get for the shock g_3 the slope

$$\frac{dy}{dx} = \frac{2}{u_L + u_R}$$

and finally the equation $y = (2x - 1)/(u_L + u_R)$. From these equations it is easy to compute the true solution pointwise. If the solution is to be known at a point Q lying within the fan region than the characteristic connecting P and Q meets the x-axis at the point $x_{PQ} = x_P + y_p(x_Q - x_P)/(y_P - y_Q)$. According to our assumed linear boundary data distribution at $y = 0$, we have $u_Q = (u_R - u_L)x_{PQ} + u_L$. This completes the description of the true solution. In Figure 6.3 the pointwise difference between the numerical and the true solution is represented.

Basic scheme with coherence measure

We start from the basic scheme (4.13) with the steering of the dissipation function by the coherence measure, i.e. (4.15). This scheme represents sharp shock resolution but still some

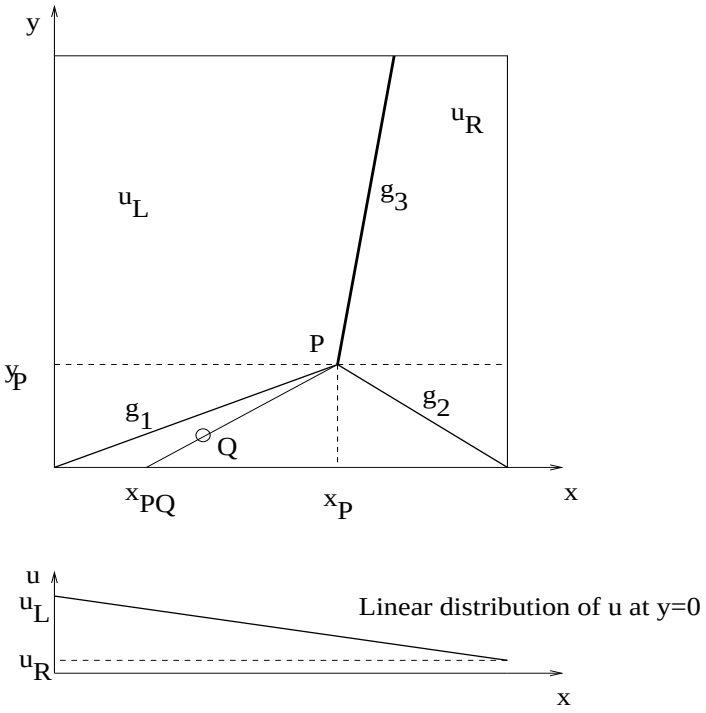

Figure 6.2: True solution of the model problem in (x, y) plane

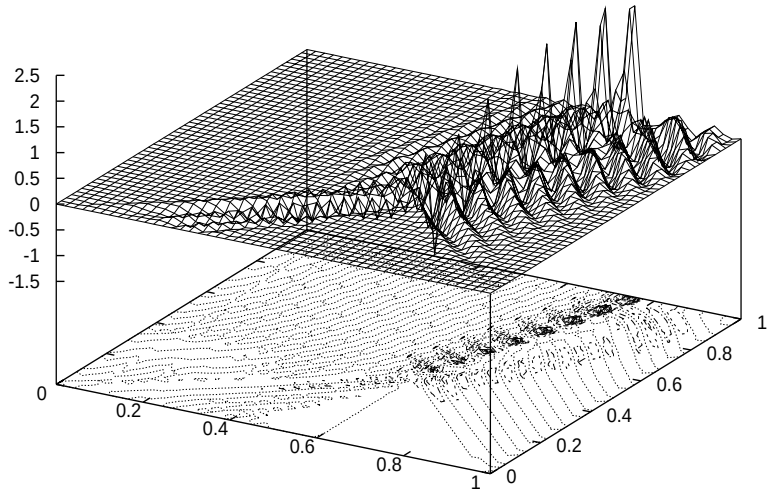

Figure 6.3: Difference between Lax-Wendroff solution and true solution

oscillations remain in the vicinity of the shock.

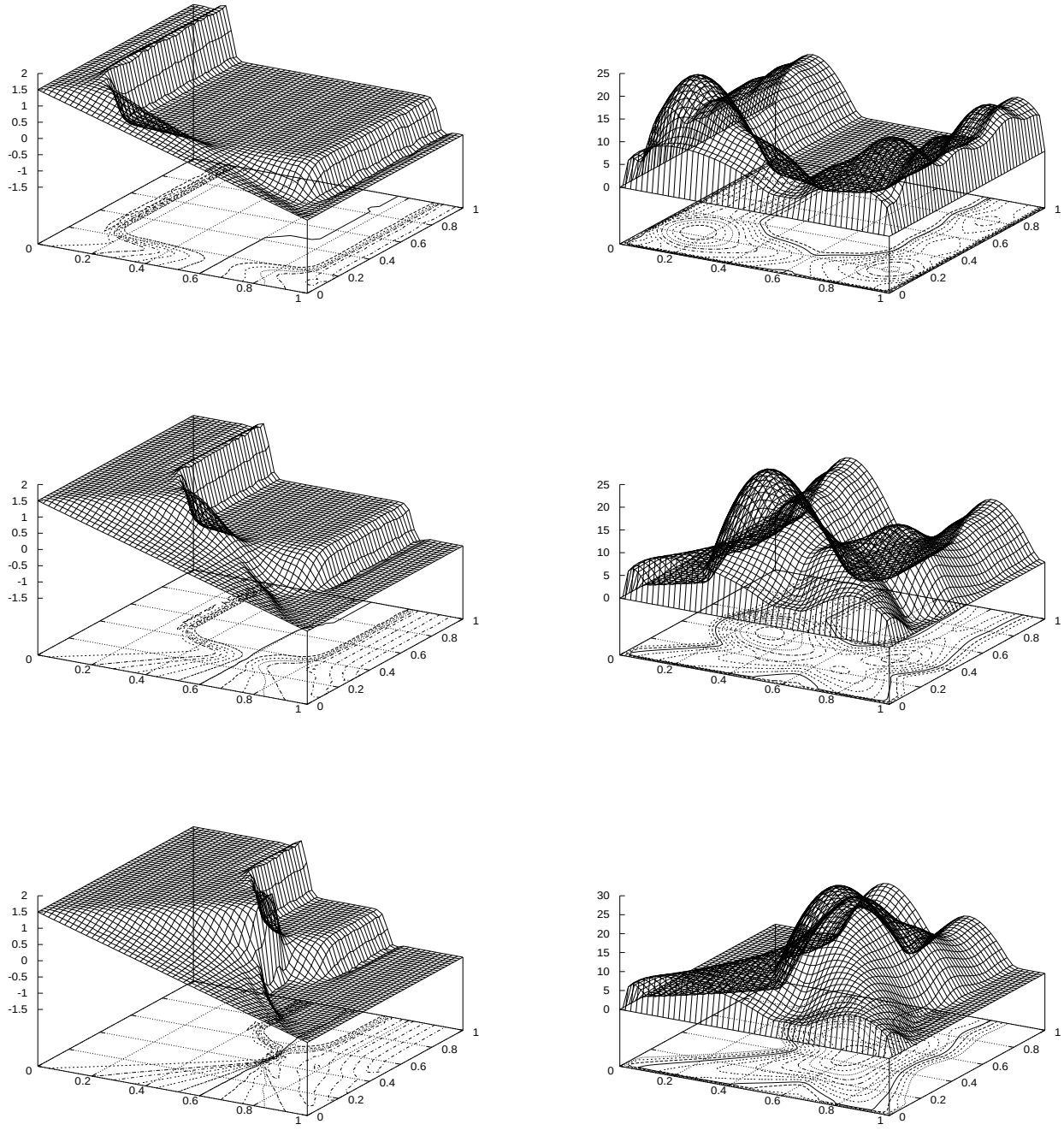

Figure 6.4: Numerical solution and coherence measure after 50, 100 and 150 time steps

Figures 6.4 and 6.5 illustrate the results after 50, 100, 150 and 200, 250 and 300 time steps respectively. In accordance with the solution the coherence measure is presented on the right. After 300 time steps the shock is formed and constantly sharpened by the diffusion step.

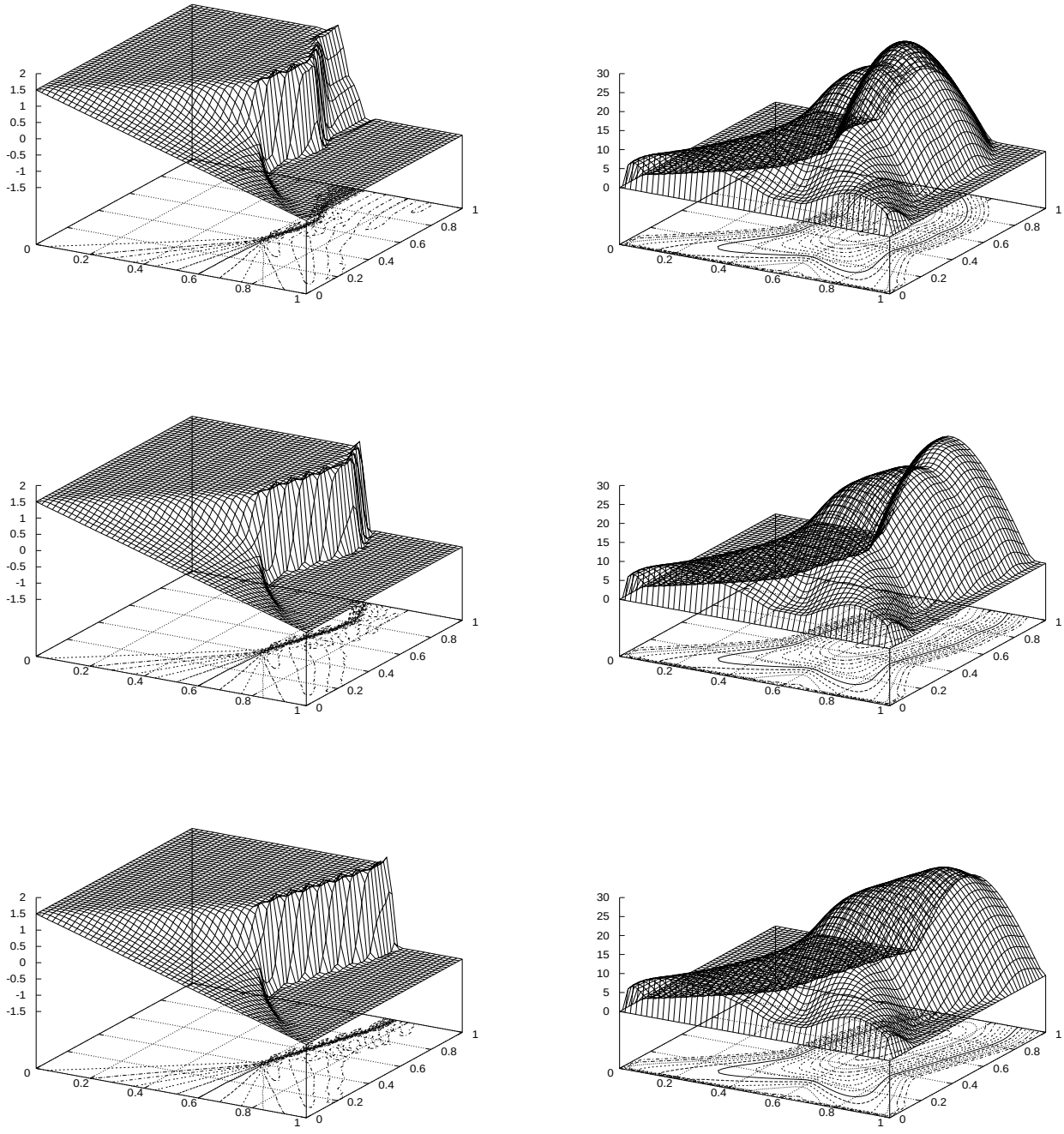

Figure 6.5: Numerical solution and coherence measure after 200, 250 and 300 time steps

Figure 6.6 shows the steady state (1000 time steps) and the corresponding coherence measure. Note that the shock is sharply resolved while there are marginal overshoots at the onset of the shock. In contrast to the result of the pure Lax-Wendroff scheme represented in Figure 6.1

we observe that the splitting scheme with the new anisotropic nonlinear artificial dissipation behaves very nicely.

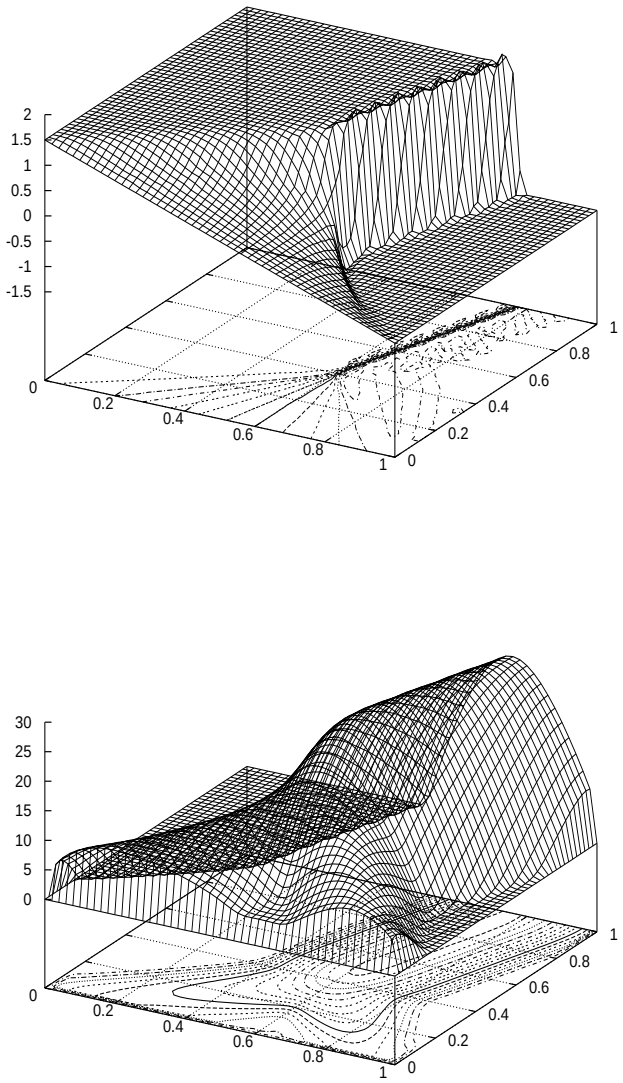

Figure 6.6: Numerical solution and coherence measure after 1000 time steps

Weighted coherence measure

In order to further reduce the small wiggles in the onset of the shock the nonlinear diffusion tensor is weighted with the derivatives of the fluxes. This procedure is quite natural if dissipation models of classical finite difference schemes are analysed. Instead of considering the structure tensor

$$\mathbf{J}_\rho(\nabla U_{i,j}) \quad = \quad \begin{bmatrix} j_{11} & j_{12} \\ j_{12} & j_{22} \end{bmatrix}$$

we therefore employ

$$\tilde{\mathbf{J}}_\rho(\nabla U_{i,j}) \;=\; \begin{bmatrix} j_{11}(f'(U_{i,j}))^2 & j_{12}f'(U_{i,j})g'(U_{i,j}) \\ j_{12}f'(U_{i,j})g'(U_{i,j}) & j_{22}(g'(U_{i,j}))^2 \end{bmatrix}.$$

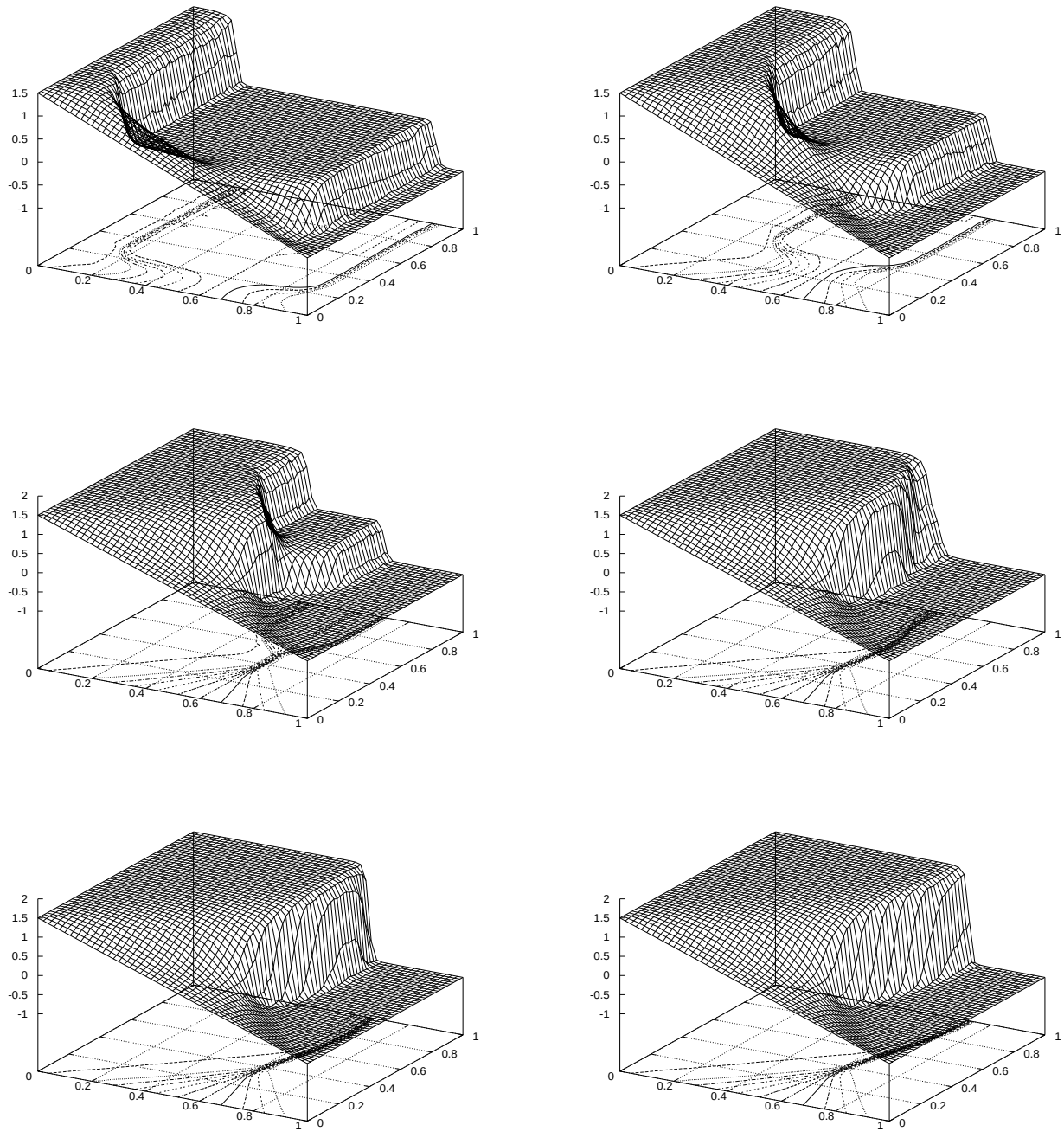

Figure 6.7: Numerical solution after 50 and 100, 150 and 200, 250 and 300 time steps

In Figure 6.7 the numerical solution for the weighted splitting scheme (4.15) at stated times is represented. Figure 6.8 depicts the steady state solution. Note that this solution exhibits not only a sharp shock transition but that it is also nearly free of any over- or undershoots.

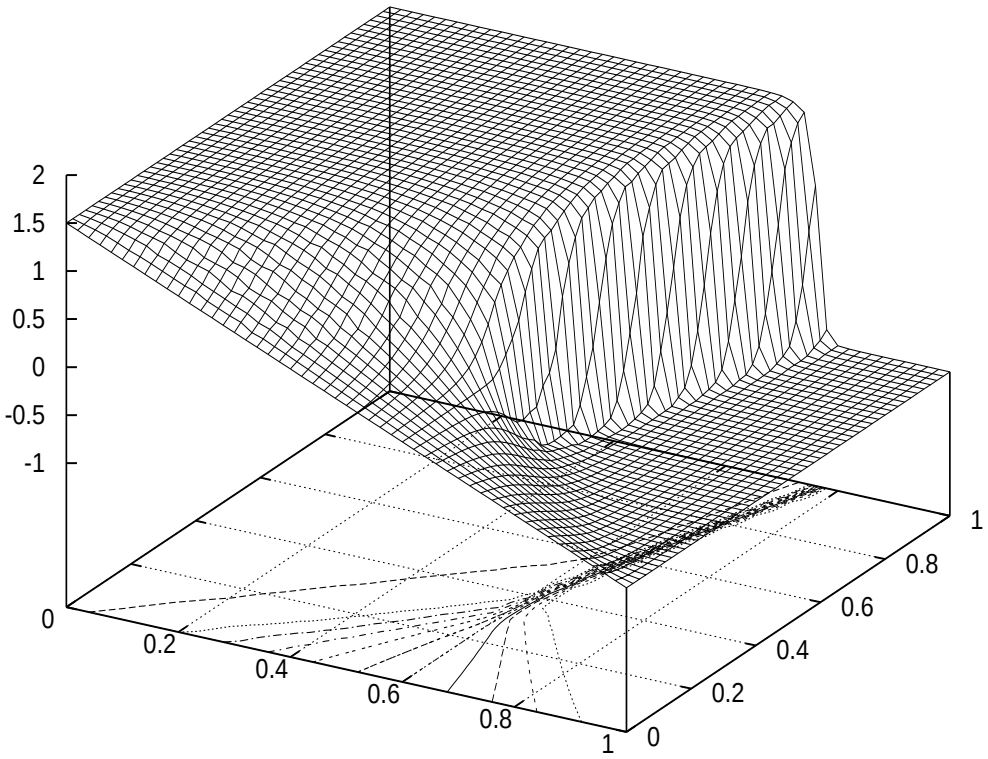

Figure 6.8: Numerical solution after 1000 time steps (steady state)

The entropy controlled blending scheme

In this section we use the scheme (4.21) based on the blending of the Lax-Wendroff type dissipation model and the entropy steered dissipation. The CFL number is 0.35. The entropy function for the indicator is chosen as $\mathsf{u} = \frac{1}{2}u^2$ with the corresponding entropy flux $\mathsf{f} = \frac{1}{3}u^3$.

Since this is a blended scheme the dissipation rate used is less than in the splitting scheme in the previous section. Thus, for the same time step the solution based on the entropy blended scheme is more advanced. Compared with the coherence based splitting scheme this solution avoids the rounding at the edges and provides sharper shock resolutions. Furthermore the computations are less expensive and the algorithm is faster. Again in Figure 6.9 the de-

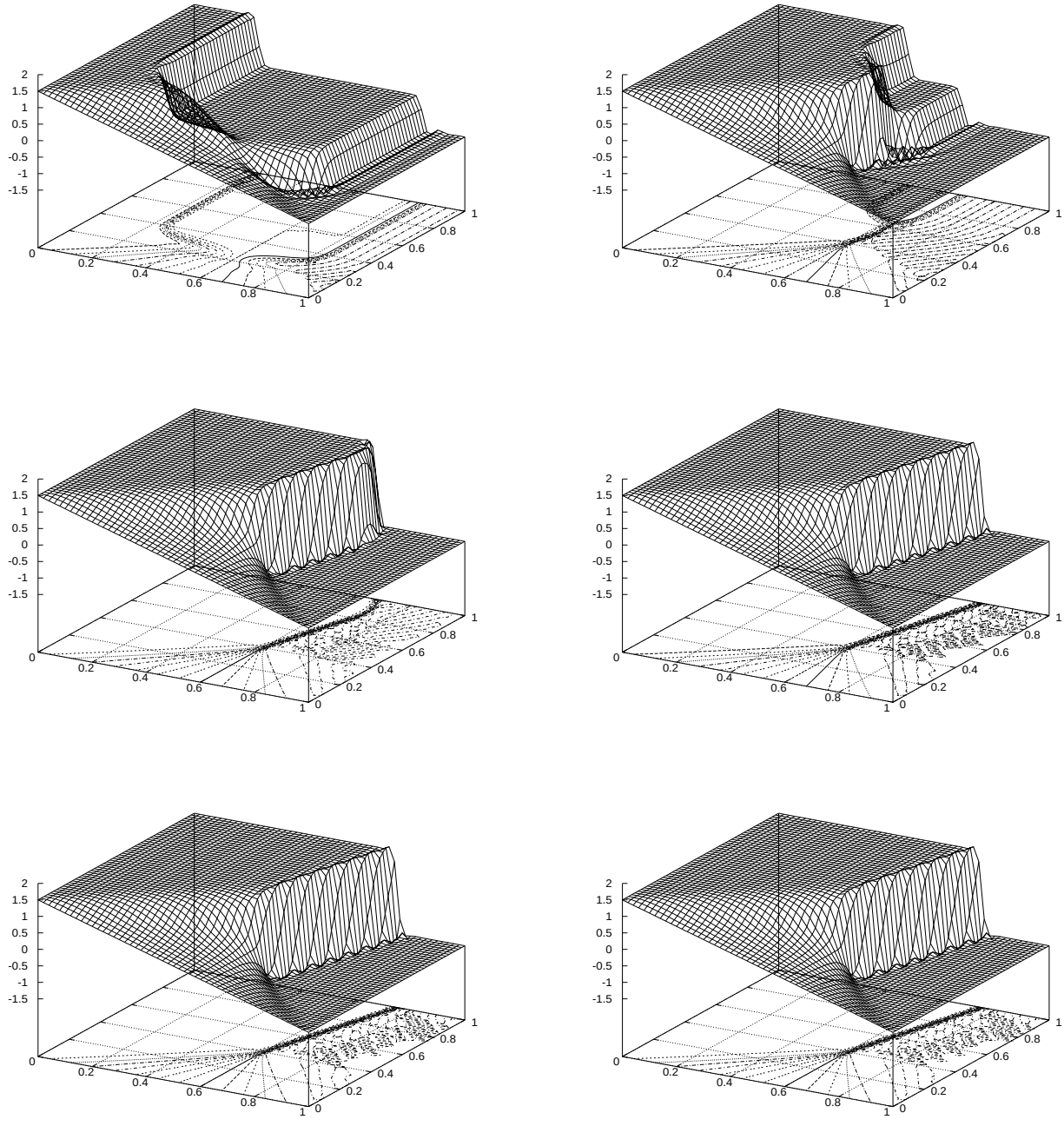

Figure 6.9: Numerical solution for the scheme (4.21) after 50 and 100, 150 and 200, 250 and 1000 time steps.

velopment in time is represented while Figure 6.10 illustrates the steady state solution after 1000 time steps.

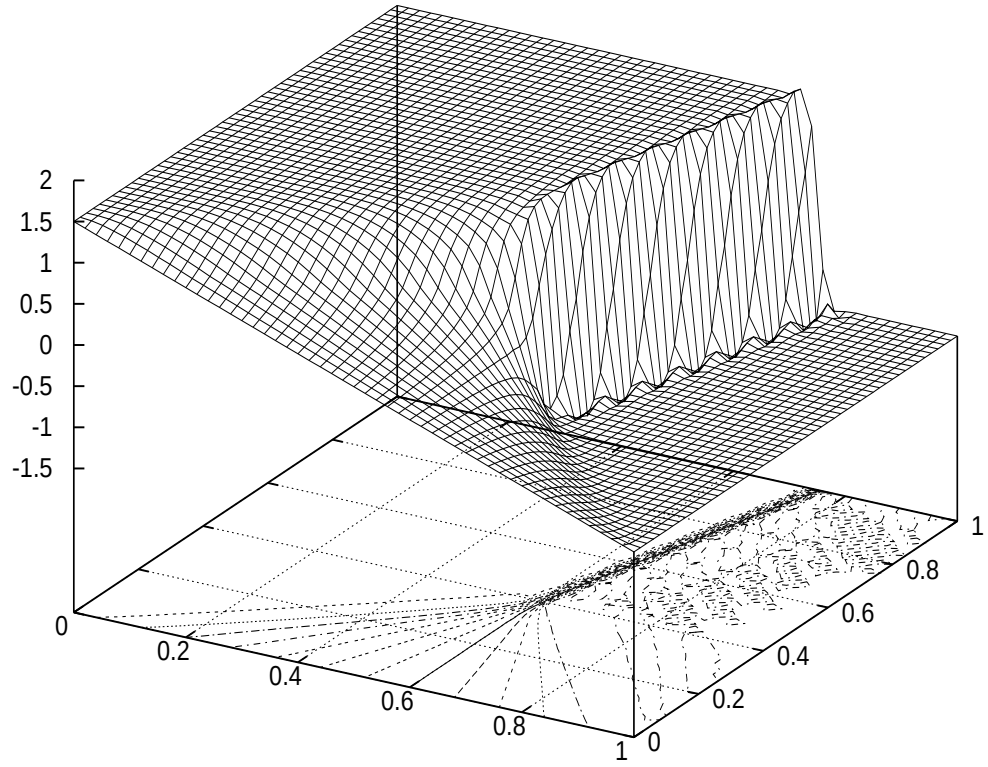

Figure 6.10: Steady state solution of the test problem applied to scheme (4.21).

Positive dissipation schemes

In this section we use the scheme (4.37) based on the unlimited LeVeque scheme and enhanced by the positive filter. Due to the excellent stabilising effect of incorporated cross fluxes we are able to choose the CFL number up to 1.

This scheme reveals by far the best shock representation with the least oscillatory behaviour. Looking at the Figures 6.11 and 6.12 one sees clearly that the discontinuity spreads over just one or two cells while oscillation occurs only at the base of the shock.

The results of the last two schemes, i.e. scheme (4.21) and scheme (4.37), show clearly that the application of entropy-steered anisotropic diffusion filters is by far the most effective way for this class of dissipation models. The entropy indicator controls the dose of dissipation and utilises it only in regions where the entropy inequality is not fulfilled.

Thus, the control of the necessary dissipation is much more effective as in the other schemes. This is confirmed by the results and the following examination concerned with the order of the schemes.

The positive scheme possesses the advantage of having a criterion to control and distribute the amount of dissipation, i.e. the positivity demand. Since it is build from the LeVeque scheme it can be used with a much higher CFL number.

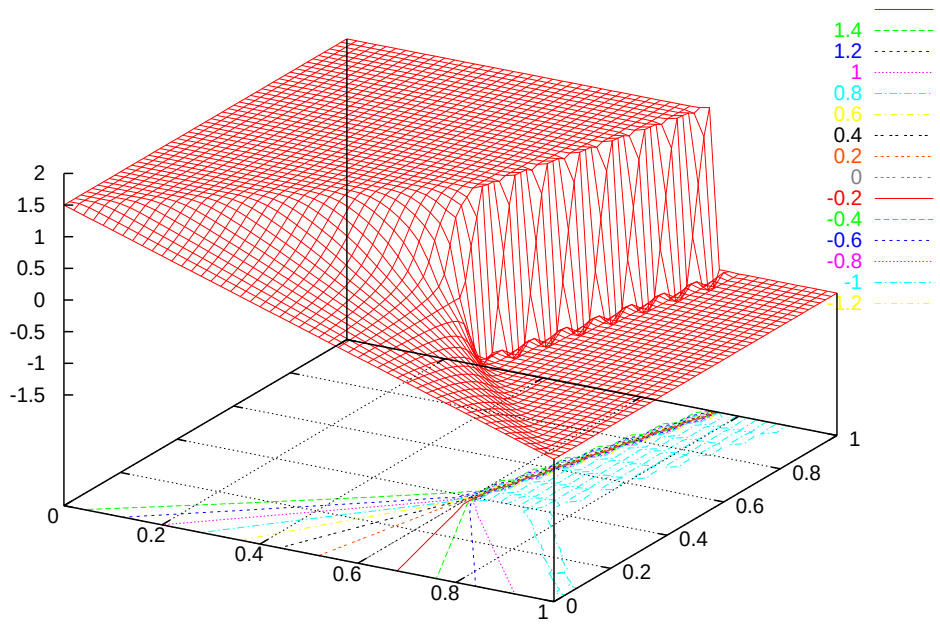

Figure 6.11: Steady state solution of the test problem

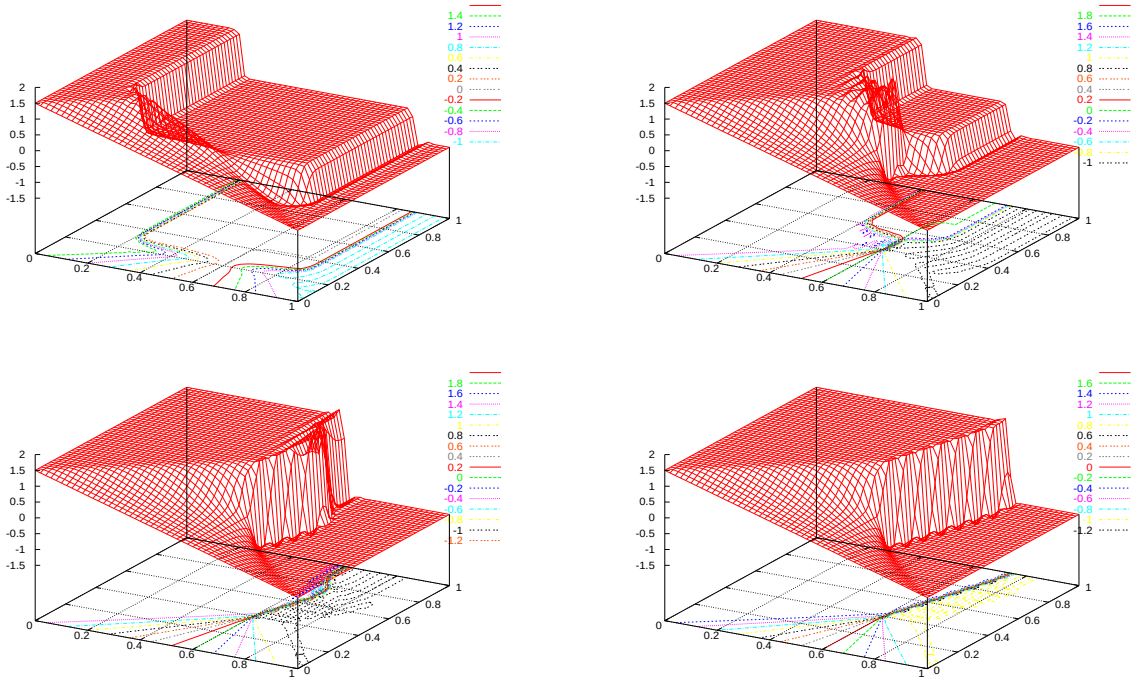

Figure 6.12: Filtered solution after 20, 40, 60 and 80 time steps

Order of convergence

Finally, we determine the order of convergence for the different schemes. Since we know the exact solution of the initial boundary value problem (6.1), the numerical error and the experimental order of convergence (EOC) can be calculated (cf. [61]).

We assume the solution u_h of a numerical scheme as the exact solution disturbed by a numerical error depending on the spatial grid size h. Thus, we write u_h as an asymptotic expansion

$$u_h \quad = \quad u + u_1 h^\alpha + \dots \quad .$$

On a coarser grid with grid size $2h$ we have the expansion

$$u_{2h} \quad = \quad u + u_1 (2h)^\alpha + \dots \quad .$$

Consequently, as a first approximation the experimental order of convergence reads as

$$\text{EOC} := \alpha = \frac{\ln\left(\frac{\|u-u_h\|_{L^1}}{\|u-u_{2h}\|_{L^1}}\right)}{\ln(\frac{1}{2})}.$$

The discrete L^1-norm is computed at time t^n by

$$\|u-u_h\|_{L^1} \quad = \quad \sum_{i,j} h^2 \left| U^n_{i,j} - \int_{y_{j-h/2}}^{y_{j+h/2}} \int_{x_{i-h/2}}^{x_{i+h/2}} u(t^n, \underline{x})\, d\underline{x} \right| .$$

The numerical errors and the experimental orders of convergence for different mesh ratios and schemes are displayed in Table 6.1.

	scheme (4.13)		scheme (4.15)		scheme (4.21)		scheme (4.37)	
h	$\|u-u_h\|_{L_1}$	EOC	$\|u-u_h\|_{L_1}$	EOC	$\|u-u_h\|_{L_1}$	EOC	$\|u-u_h\|_{L_1}$	EOC
0.08	0.168072		0.177776		0.115128		0.116748	
0.04	0.059296	1.50	0.063901	1.48	0.046160	1.31	0.039103	1.58
0.02	0.046457	0.35	0.046448	0.46	0.029755	0.63	0.023614	0.73
0.01	0.034622	0.42	0.035771	0.38	0.012953	1.20	0.009490	1.32

Table 6.1: Numerical approximation order of the schemes (4.13),(4.15),(4.21) and (4.37)

Both entropy-steered schemes perform quite well on the test case. They can be viewed as second-order schemes and are comparable to sophisticated high-order schemes using limiter function (for comparison see [61]).

The schemes using anisotropic diffusion without entropy indicators, namely the basic scheme with coherence measure (4.13) and with weighted coherence measure (4.15), have a quite dramatic loss of accuracy for small spatial grid-sizes. They have to be considered as first-order schemes.

6.2 Euler test case

As a numerical example for the inviscid Euler equations (5.1) with (5.2) we choose a test case proposed by LeVeque [72] with initial data

$$\underline{U}(0,x,y) = \begin{cases} \underline{U}_l & \text{for } \sqrt{x^2+y^2} < 0.13, \\ \underline{U}_r & \text{for } \sqrt{x^2+y^2} > 0.13, \end{cases} \tag{6.2}$$

and

$$\underline{U}_l = \begin{bmatrix} 2 \\ 0 \\ 0 \\ 15 \end{bmatrix}, \qquad \underline{U}_r = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

in primitive variables. The solution consist of a shock running outwards followed by a rarefaction wave and a contact discontinuity. A second shock moves inwards towards the centre.

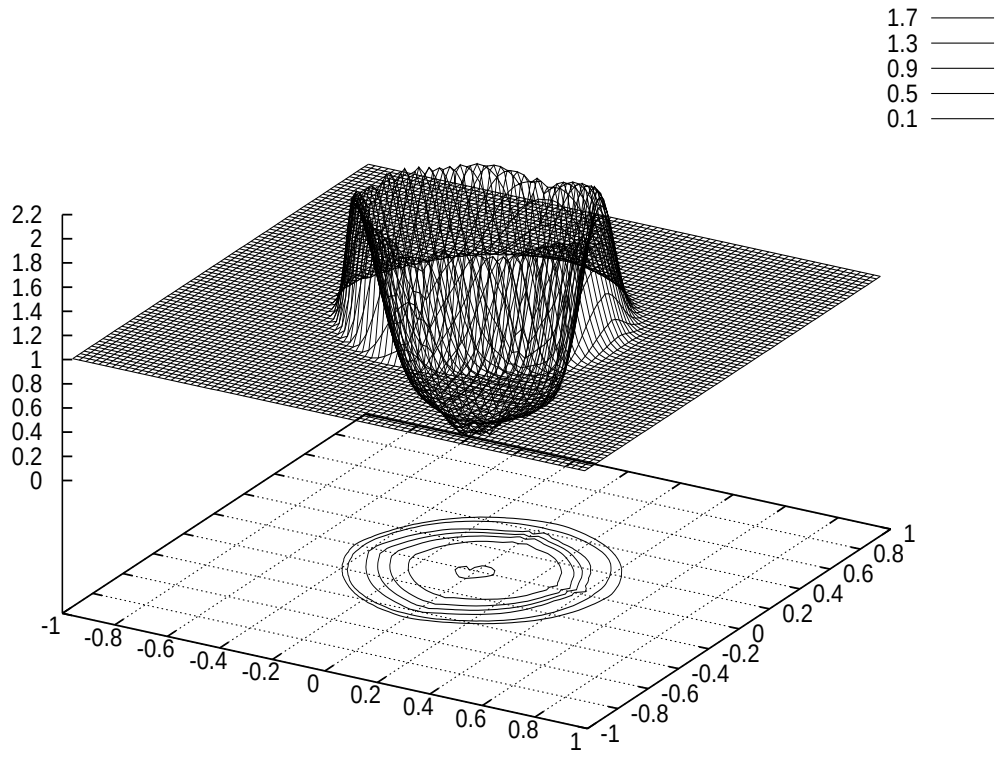

Figure 6.13: Density ρ for the test case (6.2) at time t=0.13

We use an equidistant discretisation with $\Delta x = \Delta y = h = 0.025$ and a CFL number $= 0.4$. The smoothing parameters are given with $\delta = 0.25h$ and $\nu = 0.0$ which means that no smoothing of the structure tensor takes place. The contrast parameter λ is chosen as $0.4 \max(\lambda_{1;\nu})$.

The method remains stable for this test case, which does not hold in general for a pure Lax-Wendroff scheme. Furthermore, the symmetry of the solution is maintained quite well. The solution is oscillation free and avoids strong smoothing of the shock structure. Only the amplitude of the waves differs slightly.

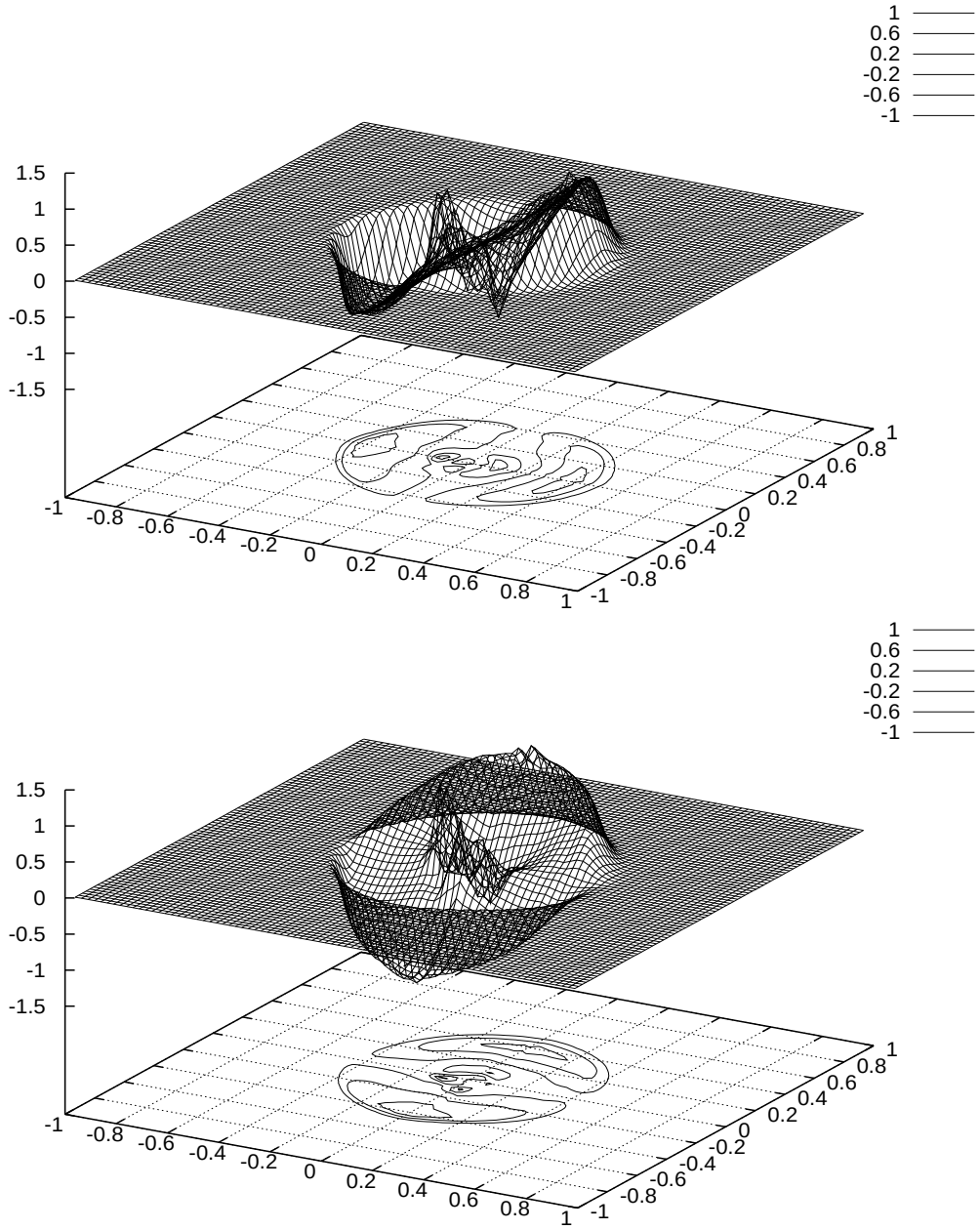

Figure 6.14: Velocity u and v for the test case (6.2) at time t=0.13

The Figures 6.13, 6.14 and 6.15 show the surface plots for the density ρ, the velocities u, v

and pressure p respectively. One sees that due to the anisotropic nature of the algorithm there are some slight asymmetries in the solution.

Figure 6.16 represents a comparison between a solution computed with the Godunov scheme and the characteristic filter algorithm for a section at $y = 0$. The Godunov solution is computed for $h = 0.01$ and CFL=0.8. In this comparison one sees clearly the good resolution of the shocks and the less dissipative behaviour of the algorithm. Special emphasise should be given to the good resolution of the inward moving shock near $x = 0.1$. One sees clearly that for a first-order method this is hard to resolve and seems to belong to the rarefaction wave. The developed characteristic filter algorithm clearly resolves the steepness of the shock and is able to distinguish between rarefaction wave and shock front.

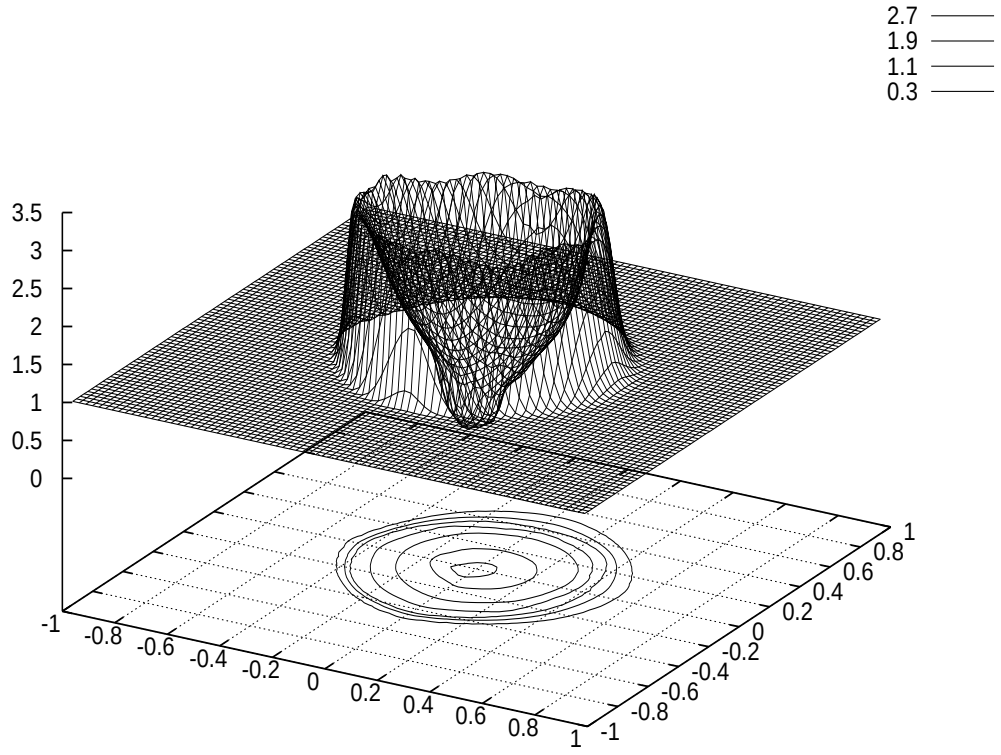

Figure 6.15: Pressure p for the test case (6.2) at time t=0.13

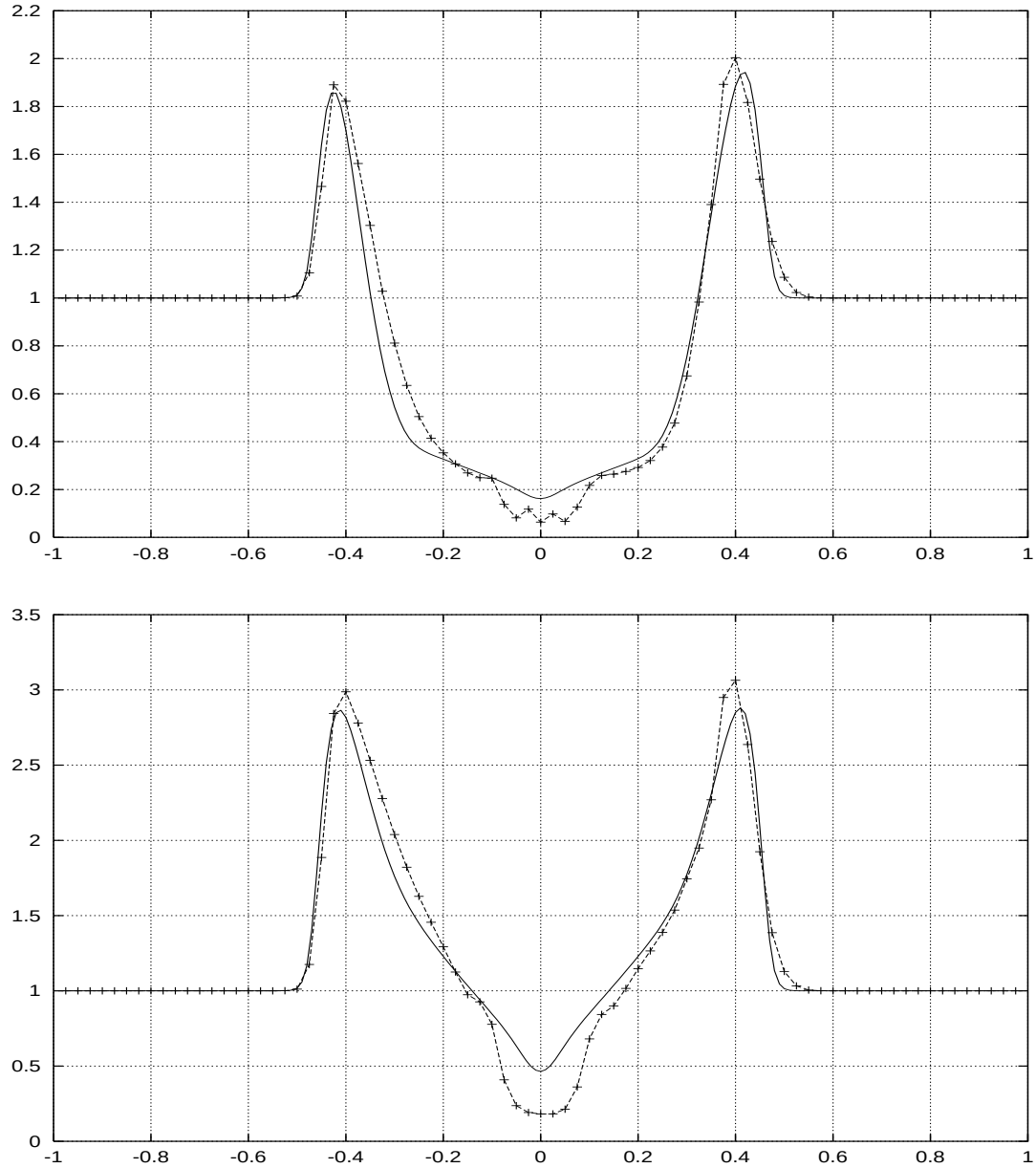

Figure 6.16: Comparison for a section at $y = 0$ between a Godunov solution (solid line) and the filtered Lax-Wendroff solution (point line) for density ρ (above) and pressure p (down) for the test case (6.2) at time t=0.13

Math is like Ophelia in Hamlet
– charming and a bit mad.

Alfred North Whitehead

7 Conclusions and perspectives

This work integrates nonlinear discrete dissipation models – namely anisotropic diffusion filters – into the field of the numerical approximation for conservation laws is accomplished. Such filters were originally designed by Weickert and had so far only be used in image processing.

The algorithms developed here on the basis of this ideas offer a new approach to data-dependent high-order methods for the numerical treatment of conservation laws since they provide a new tool to deal with nonlinear instabilities arising from highly accurate schemes. This goal has been sought for years in order to combine the fundamental requirements of accuracy and monotonicity for methods used in this field. This work describes a new direction in which we might have to follow this path.

Thus, new nonlinear anisotropic dissipation models are developed in order to stabilise high-order accurate schemes and damping the oscillatory Gibbs phenomenon in the vicinity of discontinuities. These filters are multidimensional or direction-dependent in the way that they choose the diffusion in dependence on the orientation of the discontinuity. In order to detect the orientation of the shock a discrete data analysis is performed.

For this reason an entropy indicator is developed. This detector is build from discrete cell entropy inequalities which are presented in this work. They are based on a simple choice of the numerical entropy flux according to the numerical flux of the used scheme proposed in this thesis. We prove that this choice is consistent with the classical choice for the numerical entropy flux proposed by Crandall and Majda.

From this definition of the numerical entropy flux discrete we derive dissipation models. These can be seen as equivalent in the limit with the classical Roe- or Murman-Courant-Isaacson-Rees scheme. Consequently it can be shown that the derived formulation possesses the properties of an E-scheme.

Second-order accurate methods, namely the Lax-Wendroff scheme, are stabilised using the anisotropic dissipation filters combined with the shock detector. A new class of numerical schemes with nonlinear anisotropic dissipation models are derived. These algorithms nearly reduce all oscillations while the sharp shock resolution of the high-order scheme is maintained.

Furthermore, emanating from the positivity requirement an alternative scheme is developed which is founded on the wave propagation algorithms developed by LeVeque. The corre-

sponding dissipation model can be considered as a member of the class of anisotropic diffusion filters. The developed positive diffusion filter, based on an entropy steered coefficient correction, is nearly oscillation-free and comparable to the application of limiter functions.

The numerical results show clearly that there is a strong use for a careful control of the diffusion filters. The naive integration of the anisotropic diffusion filters behaves quite well and reduce the oscillations. However, the numerical tests show they are only first-order accurate. In contrast, the entropy-steered diffusion models behave quite well. This confirms the need for a data analysis tailored to the needs of conservation laws. These schemes can be regarded as second-order accurate.

The class of anisotropic diffusion filters is extended to systems of conservation laws, namely the Euler equations. This is done by a characteristic splitting where the filter algorithm is applied to each characteristic field separately. Thus, it is shown that the extension of the class of anisotropic dissipation filter to systems of equations is possible in principle.

In the future emphasis should be given to progress in analytical results. Due to the nonlinear nature of the filter a detailed analysis is quite difficult and very hard to obtain. But at least for the case of scalar equations this has to be a major topic of future work.

In addition the generalisation to systems should be examined. We have shown that this task is basically possible but it should be feasible to produce much better results. Yet, due to the strong nonlinearities in this case this task is highly nontrivial and needs a deep understanding of the developed algorithms. This work marks the starting point in this field and leaves this task for ongoing research.

Bibliography

[1] L. Alvarez, F. Guichard, P.-L. Lions, and J.M. Morel. Axioms and fundamental equations in image processing. *Arch. Rational, Mech. Anal.*, 123:199–257, 1993.

[2] L. Alvarez, P.-L. Lions, and J.-M. Morel. Image selective smoothing and edge detection by nonlinear diffusion II. *SIAM J. Numer. Anal.*, 29:845–866, 1992.

[3] L. Alvarez and L. Mazorra. Signal and image restoration using shock filters and anisotropic diffusion. *SIAM J. Numer. Anal.*, 31(2):590–605, 1994.

[4] R. Ansorge and Th. Sonar. Informationsverlust, abstrakte Entropie und die numerische Beschreibung des zweiten Hauptsatzes der Thermodynamik. *ZAMM*, 77(11):803–821, 1997.

[5] S. Antman. The equations for large vibrations of strings. *Amer. Math. Monthly*, 87:359–370, 1980.

[6] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing – Partial Differential Equations and the Calculus of Variations.* Applied Mathematical Science Vol. 147. Springer, Berlin Heidelberg New York, 2002.

[7] H. Bateman. Some recent researches on the motion of fluids. *Monthly Weather Review*, 43:163–170, 1915.

[8] J. P. Boris and D. L. Book. Flux corrected transport: I. SHASTA, a fluid transport algorithm that works. *J. Comp. Phys.*, 11:38–69, 1973.

[9] J. P. Boris and D. L. Book. Solution of the continuity equation by the method of flux corrected transport. *J. Comp. Phys.*, 16:85–129, 1976.

[10] A. Bürgel, Th. Grahs, and Th. Sonar. From continous recovery to discrete filtering in numerical approximations of conservation laws. *Applied Numerical Mathematics*, 42:47–60, 2002.

[11] A. Bürgel and Th. Sonar. Discrete filtering of numerical solutions to hyperbolic conservation laws. *Int. J. Num. Meth. Fluids*, 40:263–271, 2002.

[12] J. Burgers. Application of a model system to illustrate some points of the statistical theory of free turbulence. *Neder. Akad. Wetensh. Proc.*, 43:2–12, 1940.

[13] S. Z. Burstein. Finite-difference calculations for hydrodynamic flows containing discontinuities. *J. Comp. Phys.*, 2:198–222, 1967.

[14] F. Catté, P.-L. Lions, J.M. Morel, and T. Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Num. Anal.*, 29(1):182–193, 1992.

[15] F. Coquel and P. LeFloch. Convergence of finite difference schemes for conservation laws in several space dimensions: The corrected antidiffusive flux approach. *Math. Comp.*, 57:169–210, 1991.

[16] R. Courant, K. O. Friedrichs, and H. Lewy. Über die partiellen Differenzengleichungen der mathematischen Physik. *Math. Ann.*, 100:32–74, 1928.

[17] R. Courant and D. Hilbert. *Methods of Mathematical Physics.* Volume II. Springer, Berlin Heidelberg New York, 1953.

[18] R. Courant, E. Isaacson, and M. Rees. On the solution of nonlinear hyperbolic differential equations by finite differences. *Comm. Pure Appl. Math.*, 5:243–255, 1952.

[19] M. G. Crandall and A. Majda. The method of fractional steps for conservation laws. *Numer. Math.*, 34:285–314, 1980.

[20] M. G. Crandall and A. Majda. Monotone difference approximations for scalar conservation laws. *Math. Comp.*, 34:1–21, 1980.

[21] C. M. Dafermos. *Hyperbolic conservation laws in continuum physics.* Grundlehren der mathematischen Wissenschaften 325. Springer, Berlin Heidelberg New York, 2000.

[22] S. F. Davis. A rotationally biased upwind difference scheme for the euler equations. *J. Comput. Phys.*, 56:65–92, 1984.

[23] B. Eilon, D. Gottlieb, and G. Zwas. Numerical stabilizers and computing time for second-order accurate schemes. *J. Comp. Phys.*, 9:387–397, 1972.

[24] B. Einfeldt. On Godunov-type methods for gas dynamics. *SIAM J. Numer. Anal.*, 25:294–318, 1988.

[25] B. Engquist, P. Lötstedt, and B. Sjögreen. Nonlinear filters for efficient shock computation. *Math.Comp.*, 52:509–537, 1989.

[26] B. Engquist and S. Osher. Stable and entropy satisfying approximations for transonic flow calculations. *Math. Comp.*, 34:45–75, 1980.

[27] B. Engquist and S. Osher. One-sided difference approximation for nonlinear conservation laws. *Math. Comp.*, 36:321–351, 1981.

[28] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics.* American Mathematical Society, Proviedence, Rhode Island, 1998.

[29] M. A. Förster and E. Gülch. A fast operator for detection and precise locationof distinct points, corners and centres of circular features. *Proc. ISPRS Int*, 21:232–237, 1987.

[30] K. O. Friedrichs and P. D. Lax. Systems of conservation equations with a convex extension. *Proc. Nat. Acad. Sci. U.S.A.*, 68:1686–1688, 1971.

[31] E. Godlewski and P.-A. Raviart. *Hyperbolic Systems of Conservation Laws.* Mathematiques & Applications 3/4. Ellipses, Paris, 1991.

[32] E. Godlewski and P.-A. Raviart. *Numerical Approximation of Hyperbolic Systems of Conservation Laws*. Applied Mathematical Science 118. Springer Verlag, 1996.

[33] S. K. Godunov. A finite difference method for the numerical computation of discontinous solutions of the equations of fluid dynamics. *Math. Sb.*, 47:271–306, 1959.

[34] Th. Grahs, A. Meister, and Th. Sonar. Nonlinear anisotropic artificial dissipation – characteristic filters for computation of the euler equations. In *Finite Volumes in complex applications*. Hermes, Paris, 1999.

[35] Th. Grahs, A. Meister, and Th. Sonar. Nonlinear anisotropic artificial dissipation for the computation of the euler equations based on algorithms from image processing. In *Proceedings of the International Symposium on Computational Fluid Dynamics*, pages 1216–1225. Bremen, 1999.

[36] Th. Grahs, A. Meister, and Th. Sonar. Image processing for numerical approximations of conservation laws: Nonlinear anisotropic artificial dissipation. *SIAM J. Sci. Comp.*, 23(5):1439–1455, 2002.

[37] Th. Grahs and Th. Sonar. Discrete cell entropy inequalities for scalar conservation laws. Internal report 01/15, TU Braunschweig, 2001.

[38] Th. Grahs and Th. Sonar. Data analysis and entropy steered discrete filter for the numerical treatment of conservation laws. *Int. J. Num. Meth. Fluids*, 40:353–359, 2002.

[39] Th. Grahs and Th. Sonar. Entropy controlled artificial anisotropic diffusion for the numerical solution of conservation laws based on algorithms from image processing. *Journal of Visual Communications and Image Representation*, 13:176–194, 2002.

[40] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods*. Pure and applied mathematics. John Wiley & Sons, New York, 1995.

[41] A. Harten. The artificial compression method for computation of shocks and contact discontinuities. III. self-adjusting hybrid schemes. *Math. Comp.*, 32:363–389, 1978.

[42] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comp. Phys.*, 49:357–393, 1983.

[43] A. Harten. On a class of high resolution total-variation-stable finite difference schemes. *SIAM J. Numer. Anal.*, 21:1–23, 1984.

[44] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy. Uniformly High Order Accurate Essentially Non-Oscillatory Schemes III. *J. Comp. Phys.*, 71:231–303, 1987.

[45] A. Harten and J. M. Hyman. Self-adjusting grid methods for one-dimensional hyperbolic conservation laws. *J. Comp. Phys.*, 50:235–269, 1983.

[46] A. Harten, J. M. Hyman, and P. D. Lax. On finite-difference approximations and entropy conditions for shocks. *Comm. Pure Appl. Math.*, 29:297–322, 1976.

[47] A. Harten, P. D. Lax, and B. van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservative laws. *SIAM Rev.*, 25:35–61, 1983.

[48] A. Harten and S. Osher. Uniformly high-order accurate nonoscillatory schemes I. *SIAM J. Numer. Anal.*, 24:279–309, 1986.

[49] A. Harten, S. Osher, B. Engquist, and S. R. Chakravarthy. Some Results on Uniformly High Order Accurate Essentially Nonoscillatory Schemes. *Appl. Num. Math*, 2:347–377, 1986.

[50] A. Harten and G. Zwas. Self-adjusting hybrid schemes for shock computations. *J. Comp. Phys.*, 9:568, 1979.

[51] Ch. Hirsch. *Numerical computation of internal and external flow*, volume 1. J. Wiley & Sons, 1988.

[52] Ch. Hirsch. *Numerical computation of internal and external flow*, volume 2. J. Wiley & Sons, 1990.

[53] E. Hopf. The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Comm. Pure Appl. Math.*, 3:201–230, 1950.

[54] M. Yousuff Hussaini, B. van Leer, and J. Van Rosendale. *Upwind and High-Resolution Schemes*. Springer Verlag, Berlin, Heidelberg, New York, 1997.

[55] A. Jameson. Iterative solution of transonic flows over airfoils and wings, including flows at Mach 1. *Comm. Pure Appl. Math.*, 27:283–309, 1974.

[56] A. Jameson, W. Schmidt, and E. Turkel. Numerical solutions of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes. *AIAA-paper*, 81–1259, 1981.

[57] F. John. *Partial Differential Equations*. Applied Mathematical Science Vol. 1. Springer, Berlin Heidelberg New York, 1991.

[58] S. Kichenassamy. The Perona-Malik paradox. *SIAM J. Appl. Math.*, 57(5):1328–1342, 1998.

[59] C. Klingenberg and S. Osher. Nonconvex scalar conservation laws in one and two space dimensions. In J. Ballmann and R. Jeltsch, editors, *Nonlinear Hyperbolic Equations – Theory, Computation Methods and Applications*, volume 24 of *Notes on Numerical Fluid Mechanics*. Vieweg, Braunschweig, 1989.

[60] G. Kreiss and G. Johansson. A note on the effect of numerical viscosity on solutions of conservation laws. unpublished.

[61] D. Kröner. *Numerical schemes for conservation laws*. Advances in numerical mathematics. John Wiley & Sons, Chichester & B.G. Teubner, Stuttgart, 1997.

[62] S. N. Kruzkov. First-order quasilinear equations in several independent variables. *Math. USSR Sbornik*, 10:217–243, 1970.

[63] F. Lafon and S. Osher. High order filtering methods for approximating hyperbolic systems of conservation laws. *J. Comp. Phys.*, 96:110–142, 1991.

[64] P. D. Lax. Hyperbolic systems of conservation laws ii. *Comm. Pure Appl. Math.*, 10:537–566, 1957.

[65] P. D. Lax. On the stability of difference approximations to solutions of hyperbolic equations with variable coefficients. *Comm. Pure Appl. Math.*, 14:497–520, 1961.

[66] P. D. Lax. Schock waves and entropy. In E.A. Zarantello, editor, *Contribution to Nonlinear Functional Analysis*, pages 603–634. Academic Press New York, 1971.

[67] P. D. Lax. Hyperbolic systems of conservation laws in several space variables. In K. Kasahara Y. Ohya and N. Shimakura, editors, *Current Topics in Partial Differential Equations*, pages 327–341. Kinokuniya Company Ltd., 1986.

[68] P. D. Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves*, volume 11 of *Lectures in Applied Math.* SIAM Regional Conf. Series, Philadelphia 1973.

[69] P. D. Lax and B. Wendroff. Systems of conservation laws. *Comm. Pure Appl. Math.*, 13:217–237, 1960.

[70] P. D. Lax and B. Wendroff. Difference schemes for hyperbolic equations with high order of accuracy. *Comm. Pure Appl. Math.*, 17:381–398, 1964.

[71] R. J. LeVeque. *Numerical Methods for Conservation Laws.* Birkhäuser, Basel, 1990.

[72] R. J. LeVeque. Simplified multi-dimensional flux limiter methods. In M. J. Baines and K. W. Morton, editors, *Numerical Methods for Fluid Dynamics 4*, pages 175–190. Oxford University Press, 1993.

[73] R. J. LeVeque. High-resolution conservative algorithms for advection in incompressible flow. *SIAM J. Numer. Anal.*, 33(2):627–665, 1996.

[74] H. W. Liepmann and A. Roshko. *Elements of Gasdynamics.* John Wiley & Sons, New York, 1957.

[75] T. Lindeberg. *Scale-Space Theory in Computer Vision.* Kluwer Academic Publisher, 1994.

[76] X.-D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *J. Comp. Phys.*, 115:200–212, 1994.

[77] A. Majda. *Compressible fluid flow and systems of conservation laws in several space variables.* Springer Heidelberg – Berlin – New York, 1984.

[78] A. Majda and S. Osher. Numerical viscosity and the entropy condition. *Comm. Pure Appl. Math.*, 32:797–838, 1979.

[79] J. Málek, J. Nečas, M. Rokyta, and M. Růžička. *Weak and Measure-valued Solutions to Evolutionary PDEs.* Applied Mathematics and Mathematical Computation. Chapman & Hall, London, 1996.

[80] M. L. Merriam. Smoothing and the second law. *Comp. Meth. App. Mech. Eng.*, 64(1):177–193, 1987.

[81] M. L. Merriam. An entropy-based approach to nonlinear stability. *NASA Technical Memorandum 101086*, 64(1):177–193, 1989.

[82] M. S. Mock. Systems of conservation laws of mixed type. *J. Diff. Equ.*, 37:70–88, 1980.

[83] S. Müller. Theory and numerics for conservation laws. Lecture notes, report 96, IGPM, RWTH Aachen, 1994.

[84] E. M. Murman. Analysis of embedded shock waves calculated by relaxation. *AIAA J.*, 12:626–633, 1974.

[85] S. Osher. Riemann solvers, the entropy condition, and difference approximations. *SIAM J. Numer. Anal.*, 21:217–235, 1984.

[86] S. Osher and L. I. Rudin. Feature-oriented image enhancement using shock filters. *SIAM J.Num.Anal.*, 27(4):919–940, 1990.

[87] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *J. Comp. Phys.*, 79:12–49, 1988.

[88] P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12:629–639, 1990.

[89] A. R. Rao and B. G. Schunck. Computing oriented texture fields. *CVGIP: Graphical Models and Image Processing*, 53:157–185, 1991.

[90] R. D. Richtmyer and K. W. Morton. *Difference Methods for Initial-Value Problems.* Wiley-Interscience, 1967.

[91] B. Riemann. Über die Fortpflanzung ebener Luftwellen von endlicher Schwingungsweite. In *Bernhard Riemanns gesammelte mathematische Werke und Wissenschaftlicher Nachlass*, pages 157–175. Teubner Leipzig, 1892.

[92] P. L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comp. Phys.*, 43:357–372, 1981.

[93] L. I. Rudin. *Images, numerical analysis of singularities and schock filters.* PhD thesis, California Institute of Technology, 1987.

[94] M. Schonbek. Second order conservative schemes and the entropy condition. *Math. Comp.*, 44:423–468, 1985.

[95] D. Serre. *Systems of Conservation Laws 1.* Hyperbolicity, Entropies, Shock Waves. Cambridge University Press, 1999.

[96] D. Serre. *Systems of Conservation Laws 2.* Geometric structures, oscillations and Initial-Boundary Value Problems. Cambridge University Press, 2000.

[97] Yu. I. Shokin. *The Method of Differential Approximation.* Springer-Verlag, 1983.

[98] C.-W. Shu. Total-variation-diminishing time discretization. *SIAM J. Stat. Comput.*, 9:1073–1084, 1988.

[99] K. Siddiqi, B. B. Kimia, and C.-W. Shu. Geometric shock-capturing eno schemes for subpixel interpolation, computation and curve evolution. *Graphical Models and Image Processing*, 59(5):278–301, 1997.

[100] B. Sjögreen. PhD thesis, University of Uppsala, 1988.

[101] J. Smoller. *Shock Waves and Reaction-Diffusion Equations*. Grundlehren der mathematischen Wissenschaften 258. Springer Heidelberg – Berlin – New York, 1994.

[102] Th. Sonar. Entropy production in second-order three-point schemes. *Numer. Math*, 62:371–390, 1992.

[103] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[104] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21:995–1011, 1984.

[105] E. Tadmor. The large-time behaviour of the scalar, genuinely nonlinear lax-friedrichs scheme. *Math. Comp.*, 43:353–368, 1984.

[106] E. Tadmor. Numerical viscosity and the entropy condition for conservative difference schemes. *Math. Comp.*, 43:369–381, 1984.

[107] E. Tadmor. The numerical viscosity of entropy stable schemes for systems of conservation laws i. *Math. Comp.*, 49:91–103, 1987.

[108] N. L. Trefethen. The definition of numerical analysis. *SIAM News*, 43:November, 1992.

[109] B. van Leer. Towards the ultimate conservative difference scheme.I. the quest of monotonicity. *Springer Lecture Notes in Physics*, 18:163–168, 1973.

[110] B. van Leer. Towards the ultimate conservative difference scheme.II. Monotonicity and conservation combined in a second order scheme. *J. Comp. Phys.*, 14:361–370, 1974.

[111] J. von Neumann and R. D. Richtmyer. A method for the numerical calculation of hydrodynamic shocks. *J. Appl. Phys.*, 21:232–237, 1950.

[112] E. V. Vorozhtsov and N. N. Yanenko. *Methods for the Localization of Singularities in Numerical Solutions of Gas Dynamics Problems*. Springer-Verlag, 1990.

[113] G. Warnecke. *Analytische Methoden in der Theorie der Erhaltungsgleichungen*. Teubner-Texte zur Mathematik. B.G. Teubner Stuttgart - Leipzig, 1999.

[114] G. W. Wei. Shock capturing by anisotropic oscillation reduction. Technical report, Department of Computing Science, National University of Singapure, 2002. preprint, `http://www.math.ntnu.no/conservation/ 2002/007.html`.

[115] J. Weickert. Theoretical foundations of anisotropic diffusion in image processing. *Computing Suppl.*, 11:221–236, 1996.

[116] J. Weickert. *Anisotropic Diffusion in Image Processing*. B.G. Teubner, Stuttgart, 1998.

[117] H. C. Yee. Construction of explicit and implicit symmetric tvd schemes and their application. *J. Comput. Phys.*, 68:151, 1985.

[118] H. C. Yee, N. D. Sandham, and M. J. Djomehril. Low-dissipative high-order shock-capturing methods using characteristic-based filters. *J. Comput. Phys.*, 150:199–238, 1999.

[119] Y. Zheng. *Systems of Conservation Laws – Two-Dimensional Riemann Problems.* Progress in Nonlinear Differential Equations and their Applications. Birkhäuser, Basel, 2001.

Zusammenfassung

In der vorliegenden Arbeit werden nichtlineare anisotrope Diffusionsfilter aus der Bildverarbeitung in Algorithmen zur numerischen Approximation von skalaren wie auch von Systemen von Erhaltungsgleichungen integriert. Dazu werden diese Filterverfahren mit einem numerischen Verfahren für Erhaltungsgleichungen von zweiter Ordnung Genauigkeit gekoppelt, namentlich dem Schema von Lax und Wendroff. Dieses ist ein Schema von hoher Approximationsordnung, welches jedoch unphysikalische Oszillationen im Bereich einer Unstetigkeit erzeugt.

Die resultierenden Schemata zur Diskretisierung von Erhaltungsgleichungen zeichnen sich durch ein neues, nichtlineares, anisotropes Diffusionsmodell aus, welches zur Stabilisierung von numerischen Verfahren hoher Ordnung notwendig ist. Dieses, von Weickert entwickelte, Dissipationsmodell berücksichtigt echt mehrdimensionale Überlegungen, da die Dissipationsstärke nicht in Abhängig von den Achsenrichtungen gewählt wird, sondern auf der Orientierung einer eventuell vorhandenen Unstetigkeit beruht. Dabei wird, soweit möglich, zur Stabilisierung Diffusion parallel zum Stoß induziert, um die Dissipation quer zur Unstetigkeit, welche zur Glättung und Abrundung des Stoßes führt, möglichst gering zu halten. Dies führt zu oszillationsreduzierenden Verfahren von zweiter Approximationsordnung mit scharfer Auflösung der Stoßunstetigkeiten.

Zur Steuerung des anisotropen Diffusionsmodells und dessen optimalen Einsatz für Erhaltungsgleichungen werden Indikatoren entwickelt, die auf der lokalen diskreten Entropieproduktion des numerischen Verfahrens beruhen. Dazu wird ein neuer, vereinfachter Ansatz zur Wahl des numerischen Entropieflusses angegeben. Es wird gezeigt, daß diese vereinfachte Approximation des Entropieflusses konsistent mit dem klassischen Ansatz von Crandall und Majda ist.

Daran anknüpfend wird eine diskrete Zellentropieungleichungen betrachtet, aus welcher ein adaptives Dissipationsmodell für skalare Erhaltungsgleichungen entwickelt wird. Das daraus resultierende Verfahren gehört zur Klasse der sogenannten E-Schemata, genügt also einer diskreten Entropiebedingung. Weiterhin ist die Wahl dieses Dissipationsmodells konsistent mit dem klassischen Verfahren von Roe.

Basierend auf dem entwickelten numerischen Entropiefluß und der diskreten Zellentropieungleichung wird eine Entropiesteuerung für die anisotropen Diffusionsfilter entwickelt. Daraus resultieren zwei verschiedene Ansätze zur Approximation von skalaren Erhaltungsgleichungen mit nichtlinearen, anisotropen Diffusionsmodellen, welche durch Entropieindikatoren gesteuert werden.

Weiterhin werden die anisotropen Diffusionsfilter für Systeme von Erhaltungsgleichungen, namentlich den Euler-Gleichungen der Gasdynamik, angepasst. Dazu wird die charakteristische Formulierung der Euler-Gleichungen verwendet und die Filteralgorithmen auf die jeweiligen charakteristischen Felder separat angewendet. Die resultierenden Algorithmen zeigen, daß die Anwendung dieser Filterklassen auf Systeme von Erhaltungsgleichungen prinzipiell möglich ist und gute Ergebnisse liefert.

Für alle in dieser Arbeit entwickelten numerischen Verfahren zur Approximation von Erhaltungsgleichungen werden entsprechende Testfälle berechnet. Die skalaren Verfahren führen zu einer deutlichen Reduktion der Oszillationen am Stoß, die von dem zu Grunde liegenden Lax-Wendroff-Verfahren herrühren. Die Unstetigkeiten können fast oszillationsfrei mit hoher Approximationsgüte dargestellt werden. Die numerischen Versuche zeigen deutlich die Notwendigkeit einer sorgfältigen, an Erhaltungsgleichungen adaptierten Kontrolle des nichtlinearen Dissipationsmodells. Ergeben die Verfahren mit einer direkten Integration des anisotropen Diffusionsfilter zwar gute visuelle Ergebnisse, aber nur eine Approximationsgüte von erster Ordnung, so sind die Algorithmen, welche auf einer Steuerung mittels eines Entropieindikators beruhen, von der Genauigkeit zweiter Ordnung. Diese entspricht der Approximationsordnung von aktuellen, auf Flußbegrenzung mittels sogenannter limiter beruhenden, TVD-Verfahren.

Für die Euler-Gleichungen zeigt das entwickelte Verfahren gute Ergebnisse bezüglich des angewendeten Testfalls. Die Stoßauflösung ist, verglichen mit einem monotonen Verfahren erster Ordnung, namentlich dem Godunov-Verfahren, sehr gut. Die Anisotropie des Filteralgorithmusses erzeugt jedoch kleinere Asymmetrien innerhalb der Lösung.

Lebenslauf

Zur Person

Geburtsdatum	27. Oktober 1968
Geburtsort	Wolfsburg
Familienstand	ledig

Schulausbildung

08/74 – 07/84	Grundschule, Orientierungsstufe und Realschule Vorsfelde/Wolfsburg
08/84 – 06/87	Fachgymnasium Technik der Berufsbildenden Schulen II/Wolfsburg
06/87	Abitur

Zivildienst

09/87 – 04/89	Emmaus-Altenheim/Wolfsburg

Universitäre Ausbildung

10/89 – 04/92	Studium der Mathematik an der Technischen Universität Braunschweig
04/92	Vordiplom in Mathematik (Nebenfach Theoretische Physik)
10/92 – 06/94	Doppelstudium Mathematik/Ozeanographie an der Universität Hamburg
06/94 – 08/94	Stipendiat der International Summer School, Oslo
09/94 – 09/95	Studium an der Universität Oslo
10/95 – 04/97	Studium an der Universität Hamburg
10/95	Vordiplom in Ozeanographie
09/96 – 10/96	Aufenthalt am von Karman Institute for Fluid Dynamics, Brüssel
04/97	Diplom in Mathematik (Nebenfach Theoretische Ozeanographie)
10/97 – 09/99	Promotionsstipendiat der Graduiertenförderung der Universität Hamburg

Berufliche Ausbildung

10/99 – 09/02	Wissenschaftlicher Mitarbeiter am Institut für Analysis der Technischen Universität Braunschweig

Liste der Veröffentlichungen

Zeitschriften

- Th. Grahs, Th. Sonar – Entropy controlled artificial anisotropic diffusion for the numerical solution of conservation laws based on algorithms from image processing
  Journal of Visual Communications and Image Representation 13, 176–194, 2002.
- A. Bürgel, Th. Grahs, Th. Sonar – From continuous recovery to discrete filtering in numerical approximations of conservation laws
  Applied Numerical Mathematics 42, 47–60, 2002.
- Th. Grahs, A. Meister, Th. Sonar – Image Processing for Numerical Approximations of Conservation Laws: Nonlinear anisotropic artificial dissipation
  SIAM Journal on Scientific Computing Vol. 23, No. 5, 1439–1455, 2002.
- Th. Grahs, Th. Sonar – Data analysis and entropy steered discrete filtering for the numerical treatment of conservation laws,
  International Journal on Numerical Methods in Fluids 40, 353–359, 2002.

Tagungsbeiträge

- Th. Grahs, A. Meister, Th. Sonar – Nonlinear Anisotropic Artificial Dissipation: Characteristic Filters for Computation of the Euler Equations.
  (in: R. Vilsmeier, F. Benkhaldoun, D. Hänel (eds.) - Finite Volumes for Complex Applications II, Hermes Science Publ., Paris, 297–306, 1999.)
- Th. Grahs, A. Meister, Th. Sonar – Nonlinear Anisotropic Artificial Dissipation for the Computation of the Euler Equations based on Algorithms from Image Processing
  (in: Proceedings of the International Symposium on Computational Fluid Dynamics, 1216–1225, Bremen, 1999)
- Th. Grahs, Th. Sonar – Multidimensional artificial dissipation for the numerical approximation of conservation laws, (in: H. Freistühler, G. Warnecke, (eds.) – Hyperbolic Problems: Theory, Numerics, Applications, ISNM Vol. 140, Birkhäuser, Basel, 463–472, 2001)
- Th. Grahs, Th. Sonar – Discrete nonlinear filters for the numerical treatment of conservation laws (in: R. Jeltsch (ed.) - Proceedings of the Annual meeting of GAMM 2001, Birkhäuser, Basel, in print.)
- Th. Grahs and Th. Sonar – Data Analysis and Entropy steered Discrete Filtering for the Numerical Treatment of Conservation Laws

(in: M J Baines (ed.) - Numerical Methods for Fluid Dynamics VII, ICFD, Oxford University Computing Laboratory, 321–327, 2001.)

Institutsberichte

- Th. Grahs and Th. Sonar – Discrete cell entropy inequalities for scalar conservation laws, TU Braunschweig, Institute für Mathematik, report 01/15, 2001.
- Th. Grahs and Th. Sonar – Entropy controlled artificial anisotropic diffusion for the numerical solution of conservation laws based on algorithms from image processing, TU Braunschweig, Institute für Mathematik, report 00/05, 2000.
- Th. Grahs, A. Meister, Th. Sonar – Image Processing for Numerical Approximations of Conservation Laws: Nonlinear anisotropic artificial dissipation, Hamburger Beiträge zur Angewandten Mathematik, Reihe F, Computational Fluid Dynamics and Data Analysis 8, 1998.
- Th. Grahs – Fourier-Analyse aus Mittelwerten auf unstrukturierten Gittern, Hamburger Beiträge zur Angewandten Mathematik, Reihe F, Computational Fluid Dynamics and Data Analysis 6, 1998.
- Th. Grahs – Multidimensional upwind fluctuation splitting schemes for scalar conservation laws, Stagiaire report 1996–12, von Karman Institute for Fluid Dynamics, October 1996.